Richard Bradley and H. Orii Stefansson

Desire, expectation, and invariance

Desire, Expectation, and Invariance

RICHARD BRADLEY[†] and H. ORRI STEFÁNSSON[*]

[†] London School of Economics and Political Science, `r.bradley@lse.ac.uk`,

[*] Institute for Futures Studies, `orristefansson@iffs.se`

Abstract

The Desire-as-Belief thesis (DAB) states that any rational person desires a proposition exactly to the degree that she believes or expects the proposition to be good. Many people take David Lewis to have shown the thesis to be inconsistent with Bayesian decision theory. However, as we show, Lewis's argument was based on an Invariance condition that itself is inconsistent with the (standard formulation of the) version of Bayesian decision theory that he assumed in his arguments against DAB. The aim of this paper is to explore what impact the rejection of Invariance has on the DAB thesis. Without assuming Invariance, we first refute all versions of DAB that entail that there are only two levels of goodness. We next consider two theses according to which rational desires are intimately connected to expectations of (multi-levelled) goodness, and show that these are consistent with Bayesian decision theory as long as we assume that the contents of 'value propositions' are not fixed. We explain why this conclusion is independently plausible, and show how to construct such propositions.

1. Introduction

The Desire-as-Belief thesis (DAB) holds that any rational person desires a proposition exactly to the degree that she believes the proposition to be desirable or good. David Lewis, the originator of the thesis, considered it to be a version of anti-Humeanism about motivation, since, if the thesis is true, then forming beliefs about what is good would suffice to produce the requisite desires and hence to motivate action. Lewis and others have also claimed that the thesis is important for both cognitivism and objectivism (or realism) in ethics, respectively the views that moral judgements are beliefs and that these beliefs are about objective and agent-independent features of reality.

How exactly Desire-as-Belief relates to these meta-ethical views is, however, in need of clarification. As we point out in §2, Humeans and anti-Humeans, and objectivists and subjectivists, might all want some version of the Desire-as-Belief thesis to be true. So what is at stake, in the debate over the thesis, is not these meta-ethical views. Rather, the issue is whether rationality requires a particular relationship between an agent's desire for a proposition A and her beliefs about propositions that express the value of A, where this value can be interpreted in line with all of the above mentioned meta-ethical views.

Lewis (1988, 1996) famously argued that the Desire-as-Belief thesis must be false, since it conflicts with Bayesian decision theory.[1] In §3 we reproduce Lewis's second[2] argument and show that it was based on a false assumption, namely Invariance, according to which the desirability of a proposition is independent of its truth-value. As we will explain, this assumption is not only intuitively implausible but inconsistent with the (standard formulation of the) version of Bayesian decision theory that Lewis assumed in his arguments against DAB.

The main aim of this paper is to explore what consequences the rejection of Invariance has on the debate over the DAB thesis. In §4 we provide a counterexample to (both causal and evidential versions of) the most simple formulations of DAB, according to which a rational agent desires a proposition to the extent that she *believes* the proposition to be good or desirable. According to somewhat more sophisticated versions of DAB, rational agents' desires are intimately connected to their *expectation* of goodness. In §5 we consider one such thesis, called *Desire-as-Expectation*, which both Humeans and anti-Humeans about motivation might want to accept, and which is not undermined by our counterexample. We show that despite appearances, the thesis is perfectly consistent with Bayesian decision theory. We conclude (in §6) with a comparison between

[1] John Collins 1988 similarly showed that a qualitative version of DAB is inconsistent with plausible constraints on qualitative belief revision. We will be concerned only with quantitative versions of DAB in this paper.

[2] Lewis's first argument also depends on an Invariance assumption that is refuted by our argument. See footnote 6.

Desire-as-Expectation and a generalised version of DAB that allows for multiple levels of goodness, and show that our argument for Desire-as-Expectation also serves to vindicate the generalised Desire-as-Belief thesis.

2. Why be interested in the Desire-as-Belief thesis?

In this section we explain why the Desire-as-Belief thesis is important. Lewis himself took the thesis to be a particularly attractive version of anti-Humeanism about motivation. So although he recognised that there might be other ways of formulating the anti-Humean view—and hence that a refutation of DAB was not sufficient to show anti-Humeanism to be false—he did think that the truth of DAB would entail the falsity of the Humean view on motivation. Moreover, he thought that DAB entailed (what he called) ethical objectivism, that is, the view that there are objective truths about ethical reality. We believe that Lewis was mistaken on both accounts. As we show below, at least some Humeans are (or could be) committed to the Desire-as-Belief thesis (§2.1) and so might ethical subjectivists be (§2.2). Whether the most plausible or attractive versions of Humeanism and subjectivism are consistent with DAB is another question. But the point of this section is simply that some version of these views are logically consistent with DAB being true. So the debate around the DAB is not one about the truth of these meta-ethical views. Instead, the question is what rationality can consistently require of agents capable of forming attitudes to propositions about value.

2.1 DAB is not anti-Humean

The Humean view on motivation that Lewis endorsed states that 'we are moved entirely by desire: we are disposed to do what will serve our desires according to our beliefs' (Lewis 1988, p. 323). His aim in refuting the Desire-as-Belief thesis was, as previously mentioned, to refute one version of the opposing anti-Humean view. For while he did admit that there might be other ways of formalising the latter view, he took a refutation of DAB to be strong evidence against anti-Humeanism, as he took DAB to be a very attractive thesis for anti-Humeans. The 'Anti-Humean's main thesis', Lewis says, is that there are necessary connections between people's desires and their beliefs about what is good: 'It is just impossible to have a belief about what would be good and lack the corresponding desire.' (ibid, p. 324). Hence, if someone truly believes that it would be good to help an old lady cross the street, then he would necessarily desire to do so, and would thus (Lewis thinks the anti-Humean would say) be motivated to help the lady cross the street.

Having explained how he understood the distinction between the Humean and anti-Humean view about motivation, Lewis formulated the Desire-as-Belief thesis—which, recall, states that any rational agent desires a proposition to the extent that she believes the proposition to be good—as an attractive version of anti-Humeanism. But the DAB thesis is not really anti-Humean. The thesis does not say that people in general necessarily desire what they believe to be good, nor that these are one and the same mental attitude. What it does say, is that all *rational* agents maintain a particular relationship between their desires and their beliefs about the good. This is at least how we will be interpreting the thesis, and, indeed, how the thesis must be understood for it to be at all interesting (as Weintraub 2007 points out). For if DAB were a psychological claim about ordinary people, then we wouldn't need a philosophical or decision-theoretic argument to examine its plausibility: citing ordinary psychological experience (with all its confused desires and so on) would then suffice to refute the thesis.

The above point might be worth clarifying further.[3] Even if DAB is a claim about *all* rational people, it is not, on our understanding, a claim about a *necessary* connection between people's desires and their evaluative beliefs. One might accept that the relationship between desires and beliefs postulated by the DAB is a necessary condition for rationality—meaning that it holds for all rational people—without accepting that the relationship in question is a necessary connection between desires and evaluative beliefs—which would mean that it would hold for all people (that have desires and evaluative beliefs), rational or not.

Since DAB is a claim about rationality, at least some Humeans about motivation will happily accept it. In a well-known defence of the Humean theory, Michael Smith says: 'Humeans . . . need not deny the contingent coexistence of beliefs and desires . . . nor . . . that the contingent coexistence of certain beliefs and desires is rationally required'

[3] We thank a referee for *Mind* for bringing to our attention the need for this clarification.

(Smith 1987, p. 119; see also his 1994). But that is exactly what the DAB thesis (as we understand it) states: it does not claim that any beliefs are *necessarily* accompanied by some particular desires, nor that some beliefs are identical to some desires. So it is not a claim about the constitution of desires and beliefs, nor a thesis about what motivates ordinary people. Instead, it is a thesis about *rationality*.

The above shows that even if DAB is true, Humeanism about motivation need not be false. What if DAB is false: does that mean anti-Humeanism must be false as well? No: there might still be *some* connection between peoples' desires and their beliefs about the good; just not the precise or fixed relationship postulated by DAB. (As we explain in §5, this relationship might be captured by John Broome's thesis of *Desire-as-Expectation*, even if the DAB thesis is false.)

2.2 DAB is not an objectivist thesis

In his second paper on the Desire-as-Belief thesis, Lewis claimed that the truth of DAB would vindicate objectivism in ethics. If there are value propositions, belief in which are connected to desires in the way postulated by DAB, then 'some of them presumably would be true', Lewis says, and then 'we surely would want to say that the true ones were objective truths about ethical reality' (Lewis 1994, p. 307).

Contrary to Lewis's claim, it does not follow, from the existence of value propositions, that some of them are true. They might, for instance, all be indeterminate, as Graham Oddie 2001 points out. Or they might all be false. For it is possible that all propositions of the form *A is good* are false, and so are all propositions of the form *A is not good*. Value propositions would then share a peculiar feature with sentences such as 'the present king of France is bald' (assuming that Bertrand Russell 1905 was right in his interpretation of such sentences). That is, propositions of the form *it is not the case that A is good* are true, and so are propositions of the form *it is not the case that A is not good*; but all propositions of the form *A is good* and *A is not good* are false. It would, admittedly, have rather strange implications if the DAB thesis were true while all value propositions are false: it would entail that a rational agent who knows the truth about value should not desire anything. But the point is that the existence of value propositions does not, by itself, entail that some of them are true.

More importantly, there are variants of the Desire-as-Belief thesis that will seem plausible to many ethical subjectivists and non-cognitivists. The literature on the DAB thesis has mostly interpreted the value propositions that figure in the formal statement of the thesis as expressing the claim that some proposition is objectively good. However, as will become apparent in §3 when we give a precise statement of DAB, these value propositions can just as well be interpreted as expressing that some proposition is desirable to some agent, or simply that it satisfies her desires (Oddie 2001 makes a similar point). Now consider the requirement that a rational agent desires a proposition A to the extent that she believes A will satisfy her considered desires. Many subjectivists would be happy to accept this version of DAB as a requirement of rationality. Most, if not all of us often violate this requirement since we often desire things while knowing that they won't satisfy us. But subjectivists might want to say that this is never true of an ideally *rational* person. As we shall see, Lewis's argument, if successful, also refutes this subjectivist version of DAB (as long as the other assumptions of Lewis's proof are satisfied).[4]

An argument by Oddie 1994 could nevertheless be seen as showing that the falsity of DAB would be especially unwelcome news for cognitivists and other objectivists in ethics. Suppose evaluative judgements are beliefs, as cognitivists claim. Then an ideal agent would, Oddie suggests, maintain *harmony* between her desires and her ethical beliefs, and thus desire whatever she believes to be morally good or right. In other words, the ideal agent would satisfy DAB. If in addition ethical objectivism is true, then the moral beliefs of an ideal agent would all be true, Oddie suggests, and she would, moreover, only desire that which is in fact good. So again, an ideal agent would satisfy DAB. However, if Lewis's arguments against DAB are successful, then an ideal agent cannot be both harmonious and fully rational. So although Lewis's argument doesn't

[4] One of these assumptions is that there are value propositions such that rational agents are certain neither of their truth nor falsity. As a referee for *Mind* points out, a subjectivist, who accepts a version of DAB, might reject this assumption, which would mean that Lewis loses his argument against DAB (see Oddie 2001, section 4, on this point). On such a subjectivist view, the assumption in question is that rational agents are (epistemically) uncertain about what propositions will satisfy their desires. Perhaps it can be argued that the most plausible version of subjectivism should, independently of Lewis's argument, reject this assumption. However, the point remains that *some* subjectivists might be worried about Lewis's result, since it undermines their theory.

show that either objectivism or cognitivism is false, it does create a bit of a dilemma for cognitivists and objectivists, since they must accept that ideal agents are either irrational or inharmonious.

The above dilemma also arises for many types of subjectivists however. For instance, we can, as previously mentioned, interpret the value propositions that figure in the formal statement of DAB as expressing that some proposition satisfies the desires of some agent. Therefore, if Lewis's argument against DAB is successful, then subjectivists must accept that even the most rational agents cannot maintain harmony between their desires and their beliefs about what satisfies these desires. Hence, the falsity of DAB would be no less awkward for subjectivists than it is for objectivists.

2.3 DAB is a thesis about rationality

What is at stake in the debate over Desire-as-Belief is not the status of different theories of value and human motivation. Rather, the issue is whether it is plausible that *rationality* requires that there be a fixed quantitative relationship between an agent's desire that A and her beliefs concerning propositions about the value of A, where this value can be understood in a way that is acceptable to the proponents of each of the views on meta-ethics and human motivation that we have discussed above.

On the face of it, it does seem that there should be some such relationship. Given that we can understand 'desirable' in a way that best suits our theory of morality and human nature, many subjectivists and objectivists alike, cognitivists and non-cognitivists, and also both Humeans and anti-Humeans about motivation, would, we think, want to say that a rational agent desires a proposition to the degree that she believes it to be desirable, or to the degree that she believes the proposition will satisfy her desires. The formal DAB thesis that Lewis claimed to refute is compatible with these different interpretations. So if Lewis's argument against DAB is correct, then all of these versions of the thesis are false, since it follows from it that a rational agent cannot satisfy DAB. If, instead, we want to say that an *ideal* agent desires a proposition to the degree that she believes it to be desirable, then we are forced to conclude, from Lewis's arguments, that the ideal agent is not rational. So, to sum up, a complete rejection of DAB-like theses would have implications for a much broader class of theories of value and human motivation than Lewis seems to have realised.

3. Lewis's argument against DAB is unsound

In this section we examine Lewis's argument against the Desire-as-Belief thesis and show that it relies on an assumption called *Invariance* (§3.1), before explaining why we think this assumption is false and hence Lewis's argument unsound (§3.2).

3.1 Lewis's argument requires Invariance

To state Lewis's argument against DAB, let A, B, etc., be propositional variables, understood as sets of possible worlds, and $\mathcal{W}$ be the set of all worlds. Ω is the set of all propositions—that is, the set of all subsets of $\mathcal{W}$—P a subjective probability (or credence) function from Ω into the interval [0,1] and V a desirability function. P thus measures the degrees of belief of a rational agent and V the strength of her desires. Let P_{A} be an agent's revised credence function after learning that A, which, if the agent revises her beliefs in accordance with Bayesian conditioning, is equal to $P(\cdot \mid \mathrm{A})$. Likewise let V_{A} be the agent's revised desirability function after she has learned that A. Finally, from any proposition A, we construct the *halo-proposition* Å, interpreted as a proposition about the value of A; for instance, that A is desirable or that A is good.

Lewis (1988, 1996) made a number of arguments against DAB. We will state the simplest of these, found in section 4 of his second paper on 'Desire as Belief' and for which it must be assumed that the desirability measure V is bounded by 0 and 1. First, a formal statement of the version of DAB he considered:

Thesis 1 (Desire-as-Belief (DAB))**.** *For any A and according to any rational agent:*

$$V(A) = P(\mathring{A}) \tag{1}$$

In what follows we will, from time to time, refer to this version of DAB as the *simple* DAB thesis, to distinguish it from the more complicated versions that we discuss in §5 and §6.

Lewis also assumed *Invariance*, according to which the desirability of a proposition is unaffected by whether the proposition is true or false:[5,6]

Assumption (Invariance)**.** *For any A and according to any rational agent:*

$$V_A(A) = V(A) \tag{2}$$

Together DAB and Invariance imply that $P(\text{Å}) = V(\text{A}) = V_{\text{A}}(\text{A}) = P_{\text{A}}(\text{Å})$. (The last equality holds since DAB is assumed to continue to hold after a rational agent learns that A.) In other words, A and Å are probabilistically independent:

Implication (Independence (IND))**.** *For any A and according to any rational agent:*

$$P(\mathring{A}) = P_A(\mathring{A}) = P(\mathring{A} \mid A) \tag{3}$$

Why is IND problematic? It is not hard to show that even if we start with a probability function for which such independence holds, it is not guaranteed that it will continue to hold after the agent in question revises her beliefs in accordance with Bayesian conditionalisation (an example is given in next paragraph). That is, suppose that an agent's revised partial beliefs after she has learned A, represented by the probability function P_{A}, is related to her partial beliefs *before* learning A, represented by P, by the following condition: for any proposition B, $P_{\text{A}}(\text{B}) = P(\text{B} \mid \text{A}) = P(\text{A\&B})/P(\text{A})$. Then we cannot be sure that this agent will satisfy IND both before and after such revision. Hence, given Invariance, a person cannot satisfy DAB unless she fails to update via Bayesian conditionalisation (i.e. unless she violates what we'll call BAYES). So if, as Lewis assumed, Invariance is true and BAYES is a requirement of rationality, then DAB cannot be rationally required (assuming that rationality does not make inconsistent demands).

Here is an example where IND and BAYES cannot both be satisfied. Assume that

[5] Lewis's first argument against DAB contained an Invariance assumption that was limited to maximally specific propositions (see e.g. Lewis 1988, p. 327). As should be apparent in §3.2, Jeffrey's decision theory, within which Lewis's discussion of DAB takes place, is also inconsistent with that version of the Invariance assumption (given how Jeffrey understands the desirability of the tautology).

[6] Costa, Collins and Levi's 1995 argument against DAB also relies on Invariance.

there is some proposition A such that $0 < P(\mathrm{A}), P(\mathring{\mathrm{A}}) < 1$ and $P(\mathrm{A} \vee \mathring{\mathrm{A}}) < 1$. (If we cannot make these assumptions, without undermining DAB, the thesis only holds in quite trivial cases, as Lewis points out.) This of course implies that $0 < P(\mathrm{A} \vee \mathring{\mathrm{A}})$. Hence, it should be possible for an agent to learn that $A \vee \mathring{\mathrm{A}}$ and we should have no problems with conditionalising on this proposition, using Bayesian conditioning. Moreover, suppose IND holds before the agent in question learns $A \vee \mathring{\mathrm{A}}$; that is, $0 < P(\mathring{\mathrm{A}} \mid \mathrm{A}) = P(\mathring{\mathrm{A}}) < 1$. Now the problem is that given these assumptions, when P is updated by $A \vee \mathring{\mathrm{A}}$ using Bayesian conditionalisation, IND no longer holds: This update leaves the conditional probability of Å given A unchanged, but increases the probability of Å (since $P(\mathring{\mathrm{A}}) < 1$). Hence, when $P(\mathring{\mathrm{A}}) = P(\mathring{\mathrm{A}} \mid \mathrm{A})$ and $0 < P(\mathrm{A}), P(\mathring{\mathrm{A}}), P(\mathrm{A} \vee \mathring{\mathrm{A}}) < 1$, then $P_{A \vee \mathring{\mathrm{A}}}(\mathring{\mathrm{A}}) \neq P_{A \vee \mathring{\mathrm{A}}}(\mathring{\mathrm{A}} \mid \mathrm{A})$. Less formally, if IND holds before an agent has learned $A \vee \mathring{\mathrm{A}}$, then she cannot still satisfy IND after having learned this unless she violates BAYES.

3.2 Why Invariance is false

In making his argument against the Desire-as-Belief thesis, David Lewis drew on the version of decision theory developed by Richard Jeffrey (1965). According to Jeffrey's theory, the desirability of any proposition, A, is a weighted average of the different mutually exclusive and jointly exhaustive possibilities compatible with A, where the weight on each of these possibilities, S_i, is given by $P(\mathrm{S}_i \mid \mathrm{A})$. More formally:

Desirability (Jeffrey's formula) For any $\mathrm{A} \in \Omega$ such that $P(\mathrm{A}) > 0$ and any partition $\{\mathrm{S}_i\}$ of $\mathcal{W}$:

$$V(\mathrm{A}) = \sum_{\mathrm{S}_i \in \mathcal{W}} P(\mathrm{S}_i \mid \mathrm{A}).V(\mathrm{A}\&\mathrm{S}_i) \tag{4}$$

The standard interpretation of this measure, suggested to Jeffrey by Leonard Savage (Jeffrey 1965, p. 82), is that it measures the *news value* of a proposition. In other words, the V-value of a proposition represents how much the agent in question would welcome the news of its truth. But it can be interpreted more generally as measuring the value of the difference that the proposition's truth makes, relative to the agent's expectations.

Since propositions are taken to be sets of possible worlds, there is just one tautological proposition, denoted by T , which is the set of all possible worlds. It follows from Jeffrey's formula that:

$$V(\mathrm{T}) = V(\mathrm{B} \vee \neg\mathrm{B}) = V(\mathrm{B}).P(\mathrm{B}) + V(\neg\mathrm{B}).P(\neg\mathrm{B}) \tag{5}$$

Jeffrey assumed that the desirability of the tautology was a constant; conventionally zero, which denotes the 'neutral' point in the desirability scale.[7] This is entirely natural on the news-value interpretation of desirability, since the news that the tautology is true is really no news at all. Indeed, on any interpretation of desirability in terms of the value of the difference that the truth of a proposition makes, the tautology will be neither desirable nor undesirable (hence, 'neutral') since its being true makes no difference at all (given that its truth is always already given).[8] Lewis's assumption that the desirability function is bounded by zero and one however rules out this conventional choice of zero for the tautology.[9] But the zero-normalisation is not essential here. What is important is that because the tautology is an 'empty' proposition whose truth is always certain, its desirability does not depend on the truth or falsity of any other proposition. Hence its desirability does not change as a result of learning the truth of any contingent proposition.

Now consider what happens as the agent's probability for some proposition B rises. As $P(\mathrm{B})$ approaches 1, $P(\neg\mathrm{B})$ approaches 0, and therefore $V(\neg\mathrm{B}).P(\neg\mathrm{B})$ approaches 0. Hence, since $V(\mathrm{B}).P(\mathrm{B}) + V(\neg\mathrm{B}).P(\neg\mathrm{B}) = V(\mathrm{T})$, $V(\mathrm{B})$ must approach $V(\mathrm{T})$. In other words, as B becomes more probable its desirability approaches the desirability of the tautology. Moreover, in the extreme case when the agent becomes certain that B, its desirability becomes equal to the tautology—neither desirable nor undesirable. So 'one who believes that a proposition is true cannot desire that it be true' (Jeffrey 1965, p. 63). For notice that together with Jeffrey's assumption about the tautology, equation (5) entails that:

$$V(\mathrm{T}) = V_{\mathrm{B}}(\mathrm{T}) = V_{\mathrm{B}}(\mathrm{B}).P_{\mathrm{B}}(\mathrm{B}) + V_{\mathrm{B}}(\neg\mathrm{B}).P_{\mathrm{B}}(\neg\mathrm{B}) = V_{\mathrm{B}}(\mathrm{B}) \tag{6}$$

But we cannot, of course, assume that all propositions are always considered equally

[7] Jeffrey (1977) considers relaxing this assumption. But as we will explain, we think that this assumption is sound, in particular given how 'desirability' is typically interpreted.

[8] For example, a similar argument can be made in terms of the willingness-to-give-up interpretation of desirability: If you are certain, both before and after learning B, that T is true already, then you should be willing to give up the same—that is, nothing of any value—to make T true before and after learning B.

[9] Unless all contingent propositions are equally desirable.

desirable as the tautology. Hence, the desirability of a proposition, according to an agent, is generally not independent of whether she takes it to be true, in contradiction with Invariance.

Note that the above demonstration of the inconsistency between Jeffrey's theory and Invariance also goes through if B is *maximally specific*; that is, only true at one possible world. Hence, the desirability of a possible world is not independent of whether or not it is actual.

As Jeffrey notes, the idea that a desire for A is 'neutralised' when one comes to believe that A is true, had been defended well before he wrote *The Logic of Decision*. In Plato's (360 BC) *Symposium*, for instance, Socrates claims that it is constitutive of desire that one cannot desire that which one already has:

> [E]very one who desires, desires that which he has not already, and which is future and not present, and which he has not, and is not, and of which he is in want . . . (Plato 360 BC)

If propositions are the objects of desires, as both Lewis and Jeffrey assumed, then Socrates can be read as claiming that he who desires, desires that which he does not already believe to be true. Socrates is of course not claiming that a person should desire *not* having that which he already has. Instead, the idea seems to be that by acquiring something the desire for that thing becomes neutral, just as Jeffrey's theory entails.

These implications of Jeffrey's framework accord well with intuitions about news-value. If you are almost certain that you will survive the day, then you won't be very excited to learn that you will indeed survive the day; that won't be much news to you. However, you would presumably be devastated to learn that it is not true. This is in agreement with what Jeffrey's measure entails: As a desirable proposition becomes more and more probable, the desirability of its truth approaches the neutral point in the desirability measure 'from above', while the desirability of its negation generally moves further and further down the negative part of the desirability scale.

It also accords well with other 'difference-making' theories of value. In economics, for instance, it is commonly held that one can measure the extent to which a person desires a proposition (or good) by what the person would be willing to give up to make that proposition come true (or to acquire that good). On this understanding of how to measure the strength of a person's desires, it is clear that degrees of desire do not satisfy Invariance. How much a person is willing to give up in order to make a proposition A come true, is certainly not independent of whether she takes A to be true already. Suppose the person considers A, whose truth she is uncertain about, to be desirable. In other words, she would be willing to give up at least something of value in order to make A true; let's call this something G. What about when she considers what she would be willing to give up to make A true, after having learned that A is true already? Surely, whatever she would be willing to give up to make A true after having learned that A is true already (if anything at all), should be less valuable to her than G.[10]

3.3 Lewis's argument for Invariance

Invariance, we have argued, is incompatible with standard interpretations of Jeffrey's concept of desirability. This is not enough to dismiss Lewis's argument against DAB, however, because Lewis evidently took the anti-Humean to be committed to Invariance. Indeed, his argument for Invariance suggests that he thought that the very formulation of the DAB thesis entailed a commitment to it. If this were true, our argument against Invariance would best be interpreted as a diagnosis of *why* the DAB thesis conflicts with decision theory rather than a refutation of Lewis's argument.

But why did Lewis think that an anti-Humean was committed to Invariance? In arguing for this he noted that Invariance holds for any proposition if it holds for all the maximally specific subcases making it up. But if a subcase,

> were maximally specific merely in all 'factual' aspects, . . . , then it would be no surprise if a change in belief changed our minds about how good it would be [that the subcase were true] . . . But the subcase was supposed to be maximally specific in *all* relevant aspects . . . The subcase has a maximally specific hypothesis about what would be good built right into it. So in assigning it a value, we do not need to consult our opinions about what is good. We just follow the built-in hypothesis.

[10] For further arguments against Invariance, see Stefánsson 2014.

> (Example. How good would it be if, first, pain were the sole good, and second, we were all about to be in excruciating and everlasting pain?—I have to say that this would be good, and so I value the case highly. My opinion that in fact pain is no good does not affect my valuing of the hypothetical case in which, *ex hypothesi*, pain is good. My opinion does cause me to give the case negligible credence, of course, but that is different from affecting the value.) (Lewis 1988, p. 332)

Lewis's argument can be broken into two distinct claims. First, that a maximally specific subcase entails the truth or falsity of all propositions about goodness. And second, that the desirability that a rational agent may assign to such a subcase is uniquely determined by the content of such propositions. In the simple case under consideration, for instance, the only relevant goodness propositions are the halo propositions and goodness comes in only two degrees; zero or one. So Lewis's two claims amount in this case to the assertion that there exist maximally specific propositions that entail, for any proposition A, either that $\mathring{\mathrm{A}}$ or that $\neg\mathring{\mathrm{A}}$ and that for any such proposition A, $V(\mathrm{A} \mid \mathring{\mathrm{A}}) = 1$.

Lewis meant his claims to extend to the more general case in which there are multiple levels of goodness. We discuss these more fully in later sections, but for the moment let us simply accept the existence of propositions that express the fact that the truth of A is good to degree i and denote them by $\mathring{\mathrm{A}}_i$. Then what Lewis requires more generally is the truth of what might be called the Principal Moral Principle; the claim that the desirability that A is true, conditional on A's truth being good to degree i, is just i.[11] More formally:

Principal Moral Principle: For any A and according to any rational agent:

$$V(\mathrm{A} \mid \mathring{\mathrm{A}}_i) = i \tag{7}$$

Lewis's argument is seductive, but misleading. Let us grant that the anti-Humean should accept both the existence of goodness propositions and the truth of the Principal Moral Principle. It does not follow, however, that they must accept Invariance. For

[11] See Nizzan-Rosen forthcoming for a discussion of this principle.

Lewis's argument requires not just that there are goodness propositions but that, for *every* proposition, there exists a corresponding halo proposition, including those propositions that are maximally specific both with respect to the non-evaluative facts and the goodness facts. To see why this is so, let A be any proposition to which there corresponds a halo-proposition Å. As Lewis notes, Invariance will hold for A if it holds for all the maximally specific subcases constituting A and that these subcases must include hypotheses about what would be good if they are to be truly maximally specific. For example, if being in pain has an invariant desirability then any subcase involving being in pain must contain a hypothesis about whether it is good to be in pain, that is, in the simple case, either 'it would be good to be in pain' or 'it would not be good to be in pain'. Is Invariance satisfied with respect to every such subcase? Not if changes in belief can change the desirability of any proposition that it makes true. So there must be haloed propositions corresponding to each such proposition that are either true or false at each subcase. In particular this is true of the haloed proposition Å whose truth fixes the invariant desirability of A; the subcase must specify whether it is good that it is good that A. And so on. But then it follows that there must exist what one might call *self-evaluative* propositions: propositions that make claims about their own goodness. In particular, there must exist a maximally specific proposition, MAX, which is of the form 'A is the case, B is not the case, A would be good, . . . , and it would be good if MAX were the case'. For if there were not then no subcase could be maximally specific with regard to what would be good.

It is the self-evaluative maximally specific propositions that impose the dubious Invariance condition on Jeffrey's framework. Not by any means the first instance in philosophy of self-referential propositions causing trouble! Lewis seems to to have thought that the anti-Humean must accept the existence of such self-evaluative propositions. But this is quite implausible. Acceptance of the existence of goodness propositions such as 'it would be good if it were to rain tomorrow' does not force acceptance of second-order goodness propositions such as 'it would be good if it were good that it would rain tomorrow', let alone propositions that are maximally specific with respect to higher-order goodness claims.

Nor, it should be emphasised, does formulation of the DAB thesis require the acceptance of higher order goodness propositions. For instance, Bradley and List 2009 exhibit a framework in which the DAB thesis can be satisfied but in which self-evaluative propositions do not exist. The key idea is to distinguish a set $\mathcal{W}$ of factual worlds, subsets of which are the purely factual propositions, from a set $\mathcal{V}$ of evaluative worlds, subsets of which are purely evaluative propositions, such as the halo-propositions or goodness-level propositions. Intuitively, we can think of factual worlds as capturing all the physical facts and evaluative worlds as capturing all the goodness facts about these physical facts. A basic (or maximally specific) possibility in this framework is just a pair (w, v) such that $w \in \mathcal{W}$ and $v \in \mathcal{V}$, and an extended proposition is any set of such world-pairs.

Now notice that on this construction there exists, for every factual proposition A, a corresponding halo proposition Å, but no haloed halo propositions. Nonetheless, the framework is rich enough to state a version of the simple DAB thesis adequate for both those Humeans and those anti-Humeans that want to endorse it. In particular, since the extended propositions form a Boolean algebra we can define a Jeffrey desirability function V and a probability function P on the set of them and then meaningfully require that for any *factual* proposition A, $V(\mathrm{A}) = P(\mathrm{Å})$. Invariance, on the other hand, is simply not required by the framework. There is thus a sense in which Lewis's argument for Invariance rests on a careless formal construction, one which allows for indiscriminate 'haloing' of propositions!

3.4 Concluding remarks on Invariance

We have shown both that Invariance is at odds with standard interpretations of desirability and that the DAB thesis does not require it. Our view is that it should therefore be dispensed with. But some of those interested in the DAB thesis might nonetheless be uncomfortable with abandoning Invariance. Ethical objectivists in particular might consider it unacceptable for rational belief about the goodness of a proposition to depend on how probable it is that the proposition is true. Hence, as both Weintraub 2007 and Daskal 2010 indirectly point out, there might seem to be a need to consider other possible responses to the tension between Invariance and standard interpretations of (Jeffrey) desirability. But this perception is based on a misunderstanding. Invariance fails in Jeffrey's framework because of the way desirability is cardinalised and in particular because it is normalised with respect to the tautology. But this does not imply that the goodness *ordering* of worlds varies with changes in belief. On the contrary, if we take the ordering of worlds to be objective, then we can construe desirability as a normalised measure of goodness, that coheres with the betterness ranking in the sense that for all worlds w and w' with non-zero probability, $V(w) \geq V(w')$ just in case w ranks at least as high as w' in the objective betterness ranking of worlds. So an ethical objectivist can subscribe to a DAB thesis formulated within Jeffrey's framework, despite the fact that it implies that rational belief about the goodness of some proposition A varies with changes in the probability of A, because this variation is a feature of the quantitative representation of desirability, rather than a reflection of any changes in the underlying betterness ordering.

This argument will not move the kind of objectivist who takes numerical desirabilities to be primitive, rather than representations of an underlying betterness ordering.[12] But the idea that desirability numbers are primitive seems very implausible to us, and is certainly something that most decision theorists would reject. And objectivists should not accept this idea either, at least if they take desirability (or value) to be similar to the quantities, such as length or temperature, that we find in the natural sciences.[13] For it is a widely held contention in the theory of measurement that such quantities are not primitive, but simply representations of comparative relations such as 'longer than' or 'warmer than'. In any case, it is incumbent on the objectivist who does take numerical desirabilities to be primitive to explain where they come from and how they are measured.

For those not persuaded by these arguments, there are two other possible responses to the inconsistency between Invariance and standard interpretations of Jeffrey's framework. First, one might try to detach the DAB thesis from Jeffrey's decision theory. And second, one could retain Jeffrey's framework but give it a non-standard interpretation

[12] Robbie Williams for bringing this view to our attention.

[13] Conrad Heilmann for suggesting this analogy.

that makes it reasonable for the desirability of the tautology to vary. In the rest of this paper, however, we take Invariance to be false and explore the implications for the debate over the DAB thesis.

4. A new counterexample to Desire-as-Belief

Lewis's argument against the DAB thesis has been shown to fail because it is based on an unsound premiss. This does not of course mean that the thesis is true. Indeed, in this section, we present an example that refutes all versions of what we above called the *simple* Desire-as-Belief thesis, which is the version of the thesis that has received almost all the attention in the literature on DAB, and according to which an agent desires a proposition to the extent that she *believes* the proposition to be good. The example highlights the implausibility of assuming that goodness comes only in two degrees, as the simple DAB thesis entails. We first explain the example and show how it undermines a simple *evidential* DAB thesis (§4.1). A number of authors have proposed a *causal* decision-theoretic version of this simple DAB thesis (see, for instance, Oddie 1994 and 2001, Byrne and Hájek 1997, and Williams 2010). But as we show in §4.2, our counterexample also undermines simple causal DAB theses. The conclusion of this section is that a minimal requirement on DAB theses is that they allow for multiple degrees of goodness (as Oddie has pointed out). In §5 and §6 we consider such versions of the thesis.

4.1 Counterexample to simple evidential DAB

Suppose we are sailing with our two good friends, Ann and Bob, when suddenly both of them fall overboard and find themselves in an equally difficult situation and threatened with drowning. In that situation we *fully* believe the proposition that it would be good that Ann is saved. (Call this proposition $\mathring{A}$.) Or if we can only fully believe a tautology, then we at least believe $\mathring{A}$ as (or almost as) strongly as we believe any contingent proposition. Nothing in our example hangs on treating Ann and Bob equally, but to simplify the discussion, let us suppose that our feelings for the two are identical in all relevant respects. Thus we also believe the proposition that it would be good to save

Bob (call this proposition $\mathring{\text{B}}$) as (or almost as) strongly as we believe any contingent proposition. So for us, in that situation, $P(\mathring{\text{A}}) = P(\mathring{\text{B}})$ is close to 1. To make the discussion that follows more precise, let us assume that $P(\mathring{\text{A}}) = P(\mathring{\text{B}}) = 1 - \gamma$.

In the situation we are imagining, we would also *desire* very strongly that Ann is saved (proposition A), and would desire equally strongly that Bob is saved (proposition B).[14] But we find it much more desirable—in fact about twice as desirable—that *both* Ann and Bob are saved than that only one of them is. So assuming that the status quo (i.e. what happens without an intervention) is that both of them drown, and, moreover, that the probability that one of them is saved is independent of the other being saved, we find that $V(\text{A\&B})$ is roughly twice $V(\text{A})$. But then the Desire-as-Belief thesis dictates that the probability that it is good that both Ann and Bob are saved, $P(\text{A}\mathring{\&}\text{B})$, should be close to twice $P(\mathring{\text{A}})$. But since $P(\mathring{\text{A}})$ is close to 1, $P(\text{A}\mathring{\&}\text{B})$ can never be close to twice $P(\mathring{\text{A}})$. Thus assuming that the requirements of rationality never ban what is rationally permissible, and if we take the attitudes towards Ann and Bob expressed in the above example to be rationally permissible, it seems that DAB cannot be a requirement of rationality.

To save DAB a proponent of it could argue that, contrary to appearances, the attitudes towards Ann and Bob assumed in the counterexample are in fact irrational. There are three ways she could do this. First, she can deny that $P(\mathring{\text{A}})$ is rationally permitted to be close to 1. Second, she can argue that $V(\text{A}) = V(\text{B})$ should be no greater than $(1 - \gamma)/2$. Third, she can argue that $V(\text{A\&B})$ should not be much greater than $V(\text{A}) = V(\text{B})$. Alternatively, she could can argue in a quite different vein, that our assumptions do not have a clear meaning.

Let us take each response in turn. Since we are assuming that saving both Ann and Bob is close to twice as desirable as saving one of them, the first response only works if we require that $P(\mathring{\text{A}})$ is less than 0.5. So for this response to work, we must be less certain in the proposition that it is good to save Ann (or Bob) than the proposition that a fair coin lands heads up when tossed. It is highly implausible that this is a rationality requirement on beliefs.

[14] Assuming that the probability of them being saved is roughly equal. The significance of this assumption will become clear at the end.

In fact, we can make things much worse. Suppose now that we are sailing with not just two but a number of our dear friends when suddenly all of them fall overboard. For the first response to the above counterexample to work, the credence we assign the proposition that it is good to save any one of our friends must get smaller and smaller as we increase the number of people that we imagine to have fallen overboard. But no matter how many friends we have, and how many of them we take out sailing, we would always be almost certain that it would be good to save each of them after having fallen overboard. To take an example, suppose we are sailing with six friends who all fall overboard. Then we cannot be more certain in the proposition that it would be good to save any particular friend than in the proposition that a dice shows side six when rolled! A conception of rationality that requires this seems very implausible.

The second route to saving the DAB thesis involves requiring that $V(\text{A}) \leq (1-\gamma)/2$. But however we interpret desirability, it is hard to believe that rationality requires that saving Ann (or Bob) be confined to the bottom half of the desirability scale. In any case, this requirement coupled with the Desire-as-Belief thesis implies that $P(\mathring{\text{A}}) \leq (1-\gamma)/2$. In other words, this response requires us to to be less certain in the proposition that it would be good to save Ann than in the proposition that a fair coin comes heads up if tossed. So this second response to our counterexample in the end comes down to the same as the first response and is no less implausible. And again, we can make this response even less plausible by increasing the number of people we are imagining to be in the water.

The third response consists in denying that it is permissible to judge it much more desirable to save both Ann and Bob than just one of them. Ordinary intuition (and many welfarist theories) suggests that saving both Ann and Bob would be roughly twice as desirable as saving one of them. Saving both might be *more* than twice as desirable as saving just one of them; for instance if we feel guilt for choosing to save one of them over the other, or if choosing to save one over the other creates some sort of injustice or unfairness. Or it might be slightly *less*, for instance if Bob and Ann hate each other and would be happier if the other were dead. But if we set these complementarities aside then we are left with the core judgement upon which the example is based: that the desirability of saving Ann (or Bob) is independent of whether the other is saved or not. But if this is so then it would seem to follow immediately from the assumption that saving Bob is equally desirable as saving Ann, that saving both is twice as desirable as saving one.[15]

Could there be complementarities that we are rationally required to gave weight to and which make the assumed judgement irrational? It is hard to imagine what they could be. But even if there are such complementarities they are unlikely to make enough of a difference. The problem is that to rescue DAB from our counterexample, it would need to be demonstrated that the difference between the desirability of saving Ann and of saving both must of rational necessity be very small. Suppose, for instance, that we are 90% sure that it would be good to save Ann and 95% sure that it would be good to save both Ann and Bob, so that by DAB, $V(\text{A}) = 0.9$ and $V(\text{A\&B}) = 0.95$. Then it is just slightly above 5% more desirable to save both Ann and Bob than to save one of them. That is implausible. Having saved one of our friends, we would still make a great deal of effort and be willing to risk or pay quite a lot to save the other. And it is hard to see why that would be irrational. We could of course argue about the plausibility of the exact numbers, but so long as the difference in the probability of Å and A&̊B is not great, the difference in desirability between saving both friends and just one of them must be *very* small for this last response to work; much smaller than what most people would intuitively accept.

So let us consider the final possible response to the counterexample, which works by questioning the meaningfulness of the assumption that the desirability of saving Ann and Bob is twice that of saving Ann. In the decision-theoretic framework in which DAB is stated, desirability functions are just numerical representations of preferences and only those properties of desirabilities that are analogues of properties of preferences should be considered meaningful. But the notion of ‘twice as desirable as’ fails this test, as is evidenced by the fact that a linear transformation of a desirability function (and in particular one based on different choice of zero point) will yield another desirability

[15] We assume the followign formula for conditional desirability: $V(\text{A} \mid \text{B}) = V(\text{A\&B}) - V(\text{B})$. (See Bradley 1999 for a justification of this formula.) The formula of course entails that Invariance is false. But that is not a problem for the present purposes, since the aim is to provide an argument against DAB without assuming Invariance. If the desirability of A is independent of B, we have $V(\text{A}) + V(\text{B}) = V(\text{A\&B})$, which, since $V(\text{A}) = V(\text{B})$, means that $V(\text{A\&B}) = 2V(\text{A}) = 2V(\text{B})$.

function that serves equally well to represent the underlying preference relation, but does not preserve properties such as one prospect being twice as desirable as another. So the counterexample trades on an unsustainable interpretation of desirabilities.

This objection is half-correct. It is true that a linear transformation of a desirability function does not preserve the property that we are interested in. But such transformations are ruled out by the simple version of the DAB thesis, which itself forces a particular choice of the zero and unit scaling points on the desirability function (namely, the certainly bad and the certainly good propositions). One may well object that such a choice of scale is arbitrary, but this would be a reason to object to (this version of) DAB directly. Here we assume for the purposes of the argument that the scaling of desirabilities enforced by the thesis is acceptable and then show that it leads to unacceptable conclusions.

What then does 'twice as desirable as' mean within the scope of desirabilities as regulated by the DAB thesis? Roughly this: An agent who regards prospect X as twice as desirable as prospect Y is one who is indifferent between Y being true for certain and a lottery which makes X true with probability one-half and the certainly bad prospect true otherwise. Suppose for instance that it is certainly bad that both Ann and Bob are not saved. Then it is twice as desirable that both Ann and Bob are saved as that Ann is saved, just in case the prospect of Ann being saved is just as desirable as the prospect of either both being saved or neither, with an equal probability of each.[16]

To sum up: The attitudes towards our friends Ann and Bob expressed in the above example are rationally permissible, and any attempt to save the simple Desire-as-Belief thesis in light of this counterexample forces us to have attitudes that seem counterintuitive and are certainly not rationally required. Hence, the example shows that this version of DAB must be false.

4.2 Counterexample to simple causal DAB

[16] It might be objected that this definition assumes that the agent is risk neutral. But the objection is misplaced. Lewis formulated DAB within the decision theory developed by Richard Jeffrey (1965). And in Jeffrey's framework—and, indeed, in all standard decision theories—risk attitudes are built into the desirabilities of propositions in the sense that the method for constructing a cardinal measure of desirability assumes risk neutrality with respect to desirability. Agents are not, however, assumed to be risk neutral with respect to specific goods, and agents who are, say, risk averse with respect to a particular good are modelled with a desirability function that is concave over that good.

A number of authors have argued that a version of the Desire-as-Belief thesis that is formulated in terms of causal decision theory rather than Jeffrey's evidential decision theory can withstand Lewis's criticism (see, for instance, Oddie 1994 and 2001, Byrne and Hájek 1997, and Williams 2010). As we now show, however, our counterexample undermines all simple causal versions of DAB (such as that considered by Byrne and Hájek 1997).

The main difference between causal and evidential decision theory is that the former weights consequences by probability under subjunctive supposition, or the probability of a counterfactual, where the latter weights them by conditional probability.[17] More precisely, let $P^{\Box}_{\mathrm{A}}(w_i)$ measure the probability that world w_i *would be* the case if A *were* true. Then causal decision theory prescribes maximisation of the *causal efficacy value*, U, of an 'action proposition', which is given by:[18]

$$U(\mathrm{A}) = \sum P^{\Box}_{\mathrm{A}}(w_i).V(w_i) \tag{8}$$

Correspondingly the version of DAB that the aforementioned authors propose states that:

Thesis 2 (Simple Causal DAB)**.** *For any A and according to any rational agent:*

$$U(A) = P(\mathring{A}) \tag{9}$$

Although these authors have not explicitly taken issue with Invariance, it is worth noting that this assumption is clearly not valid for causal efficacy value. This is perhaps best illustrated by the *Newcomb* decision problem (Nozick 1969) that historically provided the main motivation for causal decision theory. Recall that in this decision problem,

[17] As a referee for *Mind* has reminded us, this is not quite true of Lewis's 1981 own formulation of causal decision theory, which is stated in terms of beliefs in hypotheses about causal dependencies. Lewis himself took his formulation of causal decision theory to be equivalent to the more standard formulation in terms of counterfactuals (e.g. Gibbard and Harper 1981) or subjunctive suppositions (e.g. Joyce 1999). Whether Lewis was right or wrong in this regards is beyond the scope of this paper. But for the sake of argument, and to keep things simple, we will assume that any causal decision theory can be formulated in terms of subjunctive suppositions.

[18] Some causal decision theorists (for instance Lewis 1981) are happy to use Jeffrey's formula for desirability, but suggest we use this causal-efficacy formula for choice-worthiness. Others disagree and argue that desirability should match choice-worthiness (see e.g. Byrne and Hájek 1997). This disagreement is irrelevant to the present discussion, since for causal decision theory to save DAB, it has to be the case that whatever we call the type of value that figures in the DAB thesis, it is formalised by equation (8).

taking both boxes is evidence for, but does not cause, the emptiness of the 'black' box that could contain the larger amount of money. Now let A be the proposition that the agent takes both boxes. On the assumption that A, the black box is (almost certainly) empty. Hence, the utility of A given A is (very close to) the utility of receiving only what is in the 'opaque' box that can only contain the smaller amount of money. However, if we make the standard assumptions that causal decision theorists make when suggesting two-boxing, then the (unconditional) utility of A is far greater than the utility of receiving only what is in the opaque box. So Invariance fails for the utility measure that figures in causal decision theory, and Lewis's argument against the DAB thesis does not work against a causal version of the thesis.

But let us now see how the Simple Causal DAB (SCDAB) fares in light of the counterexample we discussed above. Any causal decision theorist would, we contend, say that the causal efficacy of A&B is roughly twice that of A only. To put it in the terminology that a causal decision theorist is most likely to relate to: The consequence of the act of successfully saving both Ann and Bob is roughly twice as valuable as the consequence of the act of successfully saving only Ann. However, a causal decision theorist will, just like anyone else, presumably be almost certain that it would be good to save Ann. But for the reasons discussed above, the above two judgements cannot both be true, if SCDAB is correct: If $U(\text{A\&B})$ is roughly twice $U(\text{A})$, then by SCDAB, $P(\mathring{\text{A\&B}})$ is close to twice $P(\mathring{\text{A}})$, which is inconsistent with the judgement that $P(\mathring{\text{A}})$ is close to one. So our counterexample undermines a simple causal version of DAB.

5. Desire-as-Expectation

In reaction to David Lewis's criticism of DAB, John Broome 1991 proposed the *Desire-as-Expectation* thesis (DAE), which avoids the criticisms that we (and Lewis) have directed against DAB. To state his thesis formally, let $\{\text{G}_i\}$ be a partition of the set of possible worlds according to how good they are, such that, for instance, *goodness-level proposition* G_j expresses the fact that the world is good to degree j. Then Broome's thesis says:

Thesis 3 (Desire-as-Expectation)**.** *For any A and according to any rational agent:*

$$V(A) = \sum_i i.P(G_i \mid A) \tag{10}$$

Broome claims that DAE is more plausible as an anti-Humean view than the one Lewis formulated. For there is no reason anti-Humeans should take there to be an equality (or identity) between desires and beliefs; instead, they should simply say that certain desires *result* from beliefs (or moral facts) alone. And if we assume that the G_i partition is determined by the beliefs of the agent we are modelling, then according to DAE, V(A) is determined by the evaluative beliefs of that agent. However, if we assume that the G_i partition is determined by facts about how good it would be if the different propositions were or would become true, then V(A) is determined by these facts.

Moreover, Broome thinks the DAE thesis should be no less acceptable to Humeans than anti-Humeans:

> Both groups can agree that one should desire something to a degree equal to the expectation of good from it. Where they differ is over what ultimately determines the goodness of a world. A Humean thinks goodness must ultimately be determined by people's desires; an Anti-Humean thinks this is not so. (Broome 1991, p. 265)

In other words, while anti-Humeans think the G_i-partition is determined by the agent's beliefs or facts about the good, Humeans take the partition to be determined by the agent's desires.

The fact that DAE allows for multiple degrees of goodness makes it deal well with the counterexample we raised in §4 to the simple DAB thesis. Recall that the simple DAB thesis entails that there are only two levels of goodness (good and not good). So we can either believe that it is good or not good that our friend Ann is saved from drowning. And when these are the two options, we are of course very confident in the former (and not at all in the latter). In other words, our credence in Å, the proposition that it is good that Ann is saved, is close to 1. But we also argued that it would be about twice as desirable to save both Ann and Bob from drowning as saving only one of them;

that is, $V(\text{A\&B})$ is roughly twice $V(\text{A})$. But given the simple DAB thesis, the latter entails that $P(\text{A}\mathring{\&}B)$ is roughly twice $P(\mathring{\text{A}})$, which is impossible since the latter is close to 1.

Once we replace the simple DAB with DAE, thus allowing for multiple levels of goodness, we avoid the above problem. There will plausibly be some positive[19] level of goodness such that we are almost certain that it is good to that degree that only Ann is saved from drowning (when the alternative is both Ann and Bob drowning). Otherwise we would have no reason for saving Ann, if we knew that Bob could not be saved. Let's call that degree i. Now consider a much higher degree of goodness, j. If we are almost certain that it is good to degree j that both Ann and Bob are saved, then it seems plausible that we would be far from certain that it is good to this high degree that only Ann is saved. And in contrast to the case where it was either good or not good that Ann is saved, we can now make this thought consistent with the intuition that we would be almost certain that it is good to *some* positive degree that only Ann is saved. But that means that DAE is consistent with the intuition that it would be much more desirable to save both of our friends from drowning than saving one of them only, even though it is good to some degree that only one of them is saved. More formally, the DAE allows for the possibility that $P(\text{G}_i \mid \text{A})$ is close to 1 and so is $P(\text{G}_j \mid \text{A\&B})$, but $V(\text{A\&B})$ is much higher than $V(\text{A})$ since j is much higher than i.

Moreover, DAE can be seen to be nothing more than a reformulation of Jeffrey's desirability equation, given the existence of a G_i-partition.[20] Since the G_i form a partition of the space of possible worlds, it follows from Jeffrey's formula that $V(\text{A}) = \sum_i V(\text{A\&G}_i).P(\text{G}_i \mid \text{A})$. But the same formula entails that:

$$\begin{aligned} V(\text{A\&G}_i) &= \sum_{w_j \in \text{A\&G}_i} V(w_j).P(w_j \mid \text{A\&G}_i) \\ &= \sum_{w_j \in \text{A\&G}_i} i.P(w_j \mid \text{A\&G}_i) \end{aligned}$$

in virtue of the fact that by definition, $V(w_j) = i$ for all $w_j \in \text{G}_i$. But $\sum_{w_j \in \text{A\&G}_i} P(w_j \mid$

[19] Recall the zero normalisation around the neutral prospect.

[20] Note that a partition with this formal structure can be interpreted in many ways (as further discussed in §5.2). So the assumption that there exists a partition with this formal structure is not to assume that there are, say, propositions about objective value. Nor does the existence of such a partition, in and of itself, refute subjectivism or Humeanism.

$\mathrm{A}\&\mathrm{G}_i) = 1$, since we are adding up the probabilities of all the cells in a partition of $\mathrm{A}\&\mathrm{G}_i$, after having conditioned on $\mathrm{A}\&\mathrm{G}_i$. So $V(\mathrm{A}\&\mathrm{G}_i) = i$. Hence, Jeffrey's formula entails DAE, given the existence of the G_i-partition.

But now we run into a very interesting problem. We have argued that DAE is implied by Jeffrey's formulation of desirability (given the existence of the G_i-partition). We have also seen that Invariance is false for desirability: The conditional desirability of A given A equals the desirability of the tautology and is typically not the same as the unconditional desirability of A. In contrast, DAE seems to entail Invariance: Learning that A does not change the value of $\sum_i i.P(\mathrm{G}_i \mid \mathrm{A})$, since that would require that $P(\mathrm{G}_i \mid \mathrm{A}) \neq P_{\mathrm{A}}(\mathrm{G}_i \mid \mathrm{A})$. And this inequality can never rationally hold, if rational agents change their beliefs by Bayesian updating.[21]

Another way to put the problem, is that it seems that DAE cannot be maintained as an agent learns new propositions. Recall that the DAE equation states that $V(\mathrm{A}) = \sum_i i.P(\mathrm{G}_i \mid \mathrm{A})$. But we know that the left hand side of this equation normally changes as an agent learns the proposition A. But the same is not true for the right hand side of this equation (assuming that agents respond to learning by Bayesian updating). Hence, if the equation is satisfied before an agent learns that A, then it cannot still be satisfied after the agent learns this proposition.

In §5.1 we offer a solution to this problem. Before doing so, we should point out that a causal version of Desire-as-Expectation is similarly entailed by causal decision theorists' concept of efficacy value. Recall that causal decision theorists in general say that the causal efficacy value of a proposition A is given by: $\sum_j V(w_j).P^{\square}_{\mathrm{A}}(w_j)$, where $P^{\square}_{\mathrm{A}}$ is meant to be a variable that can represent whatever probability causal decision theorists take to be relevant when evaluating the choice worthiness of A (e.g. objective

[21] Here is a proof of that $P(\,\cdot \mid \mathrm{A})$ does not change when we conditionalise on A: If $P_{\mathrm{A}}(\mathrm{B}) \doteq P(\mathrm{B} \mid \mathrm{A}) = P(\mathrm{A}\&\mathrm{B})/P(\mathrm{A})$, then $P_{\mathrm{A}}(\mathrm{B} \mid \mathrm{A}) = P_{\mathrm{A}}(\mathrm{B}\&\mathrm{A})/P_{\mathrm{A}}(\mathrm{A}) = [P(\mathrm{A}\&B\&\mathrm{A})/P(\mathrm{A})]/P(\mathrm{A}\&\mathrm{A})/P(\mathrm{A}) = P(\mathrm{B}\&\mathrm{A})/P(\mathrm{A}) = P(\mathrm{B} \mid \mathrm{A})$.

chance conditional on A, the image of P on A, etc.). But then we have:[22]

$$\begin{aligned} U(\mathrm{A}) &= \sum_j V(w_j).P^{\Box}_{\mathrm{A}}(w_j) \\ &= \sum_i i \sum_{w_j \in \mathrm{G}_i} P^{\Box}_{\mathrm{A}}(w_j) \\ &= \sum_i i.P^{\Box}_{\mathrm{A}}(\mathrm{G}_i) \end{aligned}$$

So causal efficacy value entails a causal Desire-as-Expectation thesis. This thesis, like the evidential one, is not undermined by our counterexample against the simple DAE, precisely because it allows for different degrees of goodness.

5.1 Conditioning with Indexical Propositions

The problem we face is the following. Jeffrey's theory implies both that the DAE thesis is true and that Invariance is false. But if agents revise their beliefs by Bayesian conditionalisation then DAE seems to imply Invariance. The aim of this section is to find a way out of this dilemma by giving a plausible explanation for why learning that A is the case changes the conditional probabilities for the G_i in such a way that the DAE equation can be sustained when updating on A.

Let's first get clear about what intuitively goes on when we change our views about some proposition. As before, let a proposition be a set of possible worlds. It is generally assumed that the content of a proposition—that is, what worlds make it up—remains fixed when an agent changes her mind. For instance, when an agent changes her probability for some proposition A, it is assumed that the worlds making up A remain the same, while their probabilities change. Similarly, when an agent changes her mind about the desirability of A, this is understood as a change in the probability distribution over the worlds within A; either from the less desirable worlds in A to the more desirable ones, or vice versa, and not as a change in the worlds constituting A.

In contrast, it is *not* correct to assume that the contents of the goodness-level propositions—the G_is—are fixed or invariant under changes in our evaluation of the

[22] The second equality holds since, by definition, $V(w_j) = i$ for all $w_j \in \mathrm{G}_i$; the third since the sum of the probabilities of all worlds in a proposition is just the probability of that proposition.

desirability (and hence, by DAE, the goodness) of other propositions. Recall that the goodness-level propositions partition the space of possible worlds. When we change our view about the goodness of some proposition A, it gets a new place within this partition. Suppose for instance that A was originally a subset of G_j and that after we change our mind or get new information about its goodness it becomes a subset of G_k. Since the content of A has not changed—it is still made up of the same worlds as before—this means that the contents of G_k and G_j must have changed—the former contains worlds it didn't contain before, whereas the latter now does not contain worlds it did contain before.

To take an example, suppose we think that it would be very bad if the Liberal Democrats won the next UK general election; call this proposition L. However, having heard the leader of the party set out its policies, we change our mind, and conclude that the party isn't as bad as we thought. Now the content of the proposition L has not changed, although the probability distribution within L has shifted (compared to before, we are now more confident that one of the better worlds in L is actual if L is true). But crucially, the situation of L within the goodness partition has changed, and now occupies one of the 'better' regions in the goodness-level partition than before.

The upshot of this is that the G_is are not strictly propositions, qua sets of possible worlds, but functions (of desirabilities) taking propositions as values. And so when an agent learns that A is true she must revise three things: Her probabilities for the possible worlds, her desirabilities for these worlds, and the contents of the proposition-valued functions G_i. In particular, when she learns that A she not only revises (upwards) the probability of any world consistent with A but also revises (in the direction of the value of the tautology) its desirability. As a result of doing so the worlds consistent with A will come to belong to different goodness-level 'propositions' than before, which in turn will imply shifts in the conditional probabilities of (some or all) G_is given A. Similarly, one should expect that for the G_i with a high (low) value of i, the conditional probability of G_i given A increases (decreases) when a person becomes more confident that A is good.

Objectivists and anti-Humeans might worry that the above suggestion builds subjectivism and Humeanism into the DAE thesis. We address this worry in §5.2. But the essential point is the same as the one made in §3.4: the goodness ordering over worlds can be fixed—and, in particular, is independent of what any agent learns—even though the numerical representation of this ordering (that is, the desirability values), for some agent, changes with her learning.

Let us now spell out more formally why our version of DAE does not entail Invariance. Recall that P_{A} and V_{A} are the agent's new probability and desirability functions after learning that A is true. Let the G_i be proposition-valued functions of desirability with $\mathrm{G}_i(V) = \{w : V(w) = i\}$ and $\mathrm{G}_i(V_{\mathrm{A}}) = \{w : V_{\mathrm{A}}(w) = i\}$. Note that $\mathrm{G}_i(V_{\mathrm{A}}) \neq \mathrm{G}_i(V)$ when Invariance fails. That is, although both express the fact that the world has goodness of level i, the worlds making them true are different. (We could say that the sentences expressing $\mathrm{G}_i(V)$ and $\mathrm{G}_i(V_{\mathrm{A}})$ have the same intensional content but different extensional contents.) Now as conditionalisation on A does not change any conditional probabilities given A, Bayesians require that $P_{\mathrm{A}}(\mathrm{G}_i(V) \mid \mathrm{A}) = P(\mathrm{G}_i(V) \mid \mathrm{A})$ and $P_{\mathrm{A}}(\mathrm{G}_i(V_{\mathrm{A}}) \mid \mathrm{A}) = P(\mathrm{G}_i(V_{\mathrm{A}}) \mid \mathrm{A})$. But because of the (possible) difference in content between $\mathrm{G}_i(V_{\mathrm{A}})$ and $\mathrm{G}_i(V)$, we have: $P(\mathrm{G}_i(V_{\mathrm{A}}) \mid \mathrm{A}) \neq P(\mathrm{G}_i(V) \mid \mathrm{A})$. And this, as we will explain in a moment, shows that our version of DAE does not entail Invariance.

So to make a version of the Desire-as-Expectation thesis compatible with the failure of Invariance, we should replace Broome's thesis with:

Thesis 4 (Desire-as-Expectation*)**.** *For any A and according to any rational agent:*

$$V(A) = \sum_i i.P(G_i(V) \mid A) \tag{11}$$

Then, contrary to what seemed to be the case, Bayesian conditioning is perfectly consistent with the fact that learning may change the conditional probability of the world being good to some degree given A. Similarly, once the three-fold effect of learning that

A is true is recognised we see that our version of DAE implies that:

$$\begin{aligned} V_{\mathrm{A}}(\mathrm{A}) &= \sum_i i.P_{\mathrm{A}}(\mathrm{G}_i(V_{\mathrm{A}}) \mid \mathrm{A}) \\ &= \sum_i i.P(\mathrm{G}_i(V_{\mathrm{A}}) \mid \mathrm{A}) \\ &\neq \sum_i i.P(\mathrm{G}_i(V) \mid \mathrm{A}) = V(\mathrm{A}) \end{aligned}$$

So Desire-as-Expectation* and Bayesian conditioning are jointly consistent with a denial of Invariance.

Our findings in this section provide support for a suggestion made by Alan Hájek and Philip Pettit 2004. They suggest that goodness is indexical in the same way we have said it must be—that is, partly a function of a person's attitudes—and they show that Lewis's argument against DAB then loses its bite. Moreover, they explain why various meta-ethical views are committed to this indexicality. There are, nevertheless, important differences between our discussion of this issue and Hájek and Pettit's. First, they accept Lewis's argument against DAB as sound and suggest an indexical DAB thesis to avoid his result. We, on the other hand, have argued that Lewis's argument is not sound, but that we nevertheless need an indexical account of goodness to save the Desire-as-Expectation thesis. Secondly, unlike Hájek and Pettit, we have shown that unless goodness is indexical, Jeffrey's decision theory leads to contradiction, since it entails both the truth of Desire-as-Expectation and the falsity of Invariance, which is inconsistent unless goodness is indexical. Finally, the indexical thesis they suggest does not assume that there are multiple degrees of goodness, and is therefore refuted by the counterexample we discussed in §4.

5.2 Anti-Humeans and objectivists can also accept DAE

We have seen that our version of Desire-as-Expectation is not only consistent with Jeffrey's decision theory, but an implication of it. But this might give rise to the worry that, contrary to Broome's claim, Desire-as-Expectation is not properly anti-Humean, nor consistent with objectivism about value, since on any defensible version of DAE, the goodness value of A may change as the agent learns that A.

This worry is, however, unnecessary. First, as mentioned in §3.4, the rejection of Invariance does not mean that the goodness *ordering* of worlds violates the corresponding invariance condition on such orderings. That ordering might be fixed by the objective good, or perhaps by some agent's beliefs. But the numerical representation of that ordering is not fixed. When an agent becomes certain that a proposition is true, the desirability (and, by DAE, the goodness value) of that proposition becomes equal to the tautology, which is conventionally assigned a zero-desirability value. But such a re-normalisation around the zero-point should still represent the underlying (and invariant) betterness ordering.

The above remarks mean that only those objectivists and anti-Humeans who think that the the numbers that are used to represent the goodness ordering—that is, the desirability values—are in some sense objective should worry about the failure of Invariance, and what it entails for DAE. But, as explained in §3.4, we cannot see why anyone would take such numbers to be in any sense objective. For the exact numbers themselves are completely arbitrary. Or, to be precise, they are arbitrary until one has chosen the scale and zero-point. So as long as the underlying betterness ordering is invariant, anti-Humeans and objectivists need not worry that the numerical representation of it violates Lewis's Invariance condition.

Another reason why the above worry is unnecessary, is that Anti-Humeans need not say that people are motivated directly through their beliefs about the good without these beliefs affecting their desires. (If they did say that, then they would have to deny that the structure of desire is captured by Jeffrey's formula.) In fact, the anti-Humean view can be characterised as precisely the idea that because rational people's beliefs about the good determine their desires, these beliefs determine what people are motivated to do. Broome, for instance, characterises the anti-Humean view thus:[23]

> Sometimes, we do what will serve the good according to our beliefs about what would be good together with our other beliefs—no desire, *other than desires which result from beliefs alone*, need enter into it. (Broome 1991, p. 266,

[23] Broome is in the quoted passage rephrasing Lewis's 1988 (p. 324) characterisation of anti-Humean. According to Lewis, anti-Humeans say that the only desires that motivate people to act are those that are *identical* with beliefs.

emphasis added)

If this is how we understand the anti-Humean view, then it is not a problem that a person's expectation of the good changes with her desires, if this change in desires is brought about by a change in beliefs. And that is, for instance, exactly what happens when expectation of good changes because a desire for a proposition has changed as a result of a change in the proposition's probability. More generally, there is no reason why an anti-Humean could not endorse our idea that the contents of the goodness propositions change with an agent's desires, but argue that what grounds such a change is (often, at least) a change in the agent's normative beliefs.

Moreover, since we can interpret the $\mathrm{G}_i(V)$ functions however we like, the Desire-as-Expectation thesis should be acceptable to subjectivists as well as objectivists about value. We could, for instance, interpret $\mathrm{G}_i(V)$ as expressing the fact that my desires are satisfied to degree i. Then DAE states that we should desire a proposition to the degree that we expect our desires to be satisfied when the proposition is true. This is something that people might want to accept irrespective of where they belong in the Humean/anti-Humean and objectivist/subjectivist divide.

6. A generalisation of Desire-as-Belief?

Instead of avoiding the counterexample we raised against the simple Desire-as-Belief thesis by switching to some version of Desire-as-Expectation, a defender of DAB might respond by generalising the original DAB thesis to a thesis that allows for multiple degrees of goodness. As we saw at the start of §5, our counterexample would not undermine such a generalisation of DAB, since what the example illustrates, is the the implausibility of assuming (as the simple DAB does) that there are just two levels of goodness. Moreover, anti-Humeans should, independently of any discussion of DAB, be skeptical of the idea that goodness comes only in two degrees. Lewis himself considered such a generalised version of DAB and made a similar argument against it to that which he made against the simple DAB thesis (Lewis 1988, p. 330–1). But this argument again assumed Invariance. Therefore, a version of DAB based on multiple levels of goodness is neither refuted by Lewis's argument nor by our counterexample. But is it consistent with the truth of Desire-as-Expectation*? We conclude the paper with an argument for their compatibility.

To state the general version of the Desire-as-Belief thesis more precisely, let $\mathring{A}_i$ be the proposition that A is good to degree i.[24] Then the thesis under consideration states that:

Thesis 5 (Generalised Desire-as-Belief (GDAB))**.** *For any A and according to any rational agent:*

$$V(A) = \sum_i i.P(\mathring{A}_i) \tag{12}$$

If (our version of) the DAE thesis is true, then GDAB entails that:

$$\sum_i i.P(\mathring{\mathrm{A}}_i) = \sum_i i.P(\mathrm{G}_i(V) \mid \mathrm{A}) \tag{13}$$

And that might seem very plausible. For one might think that, say, the probability that it is good to degree i that Ann is saved from drowning, should be equal to the conditional probability that the world is good to degree i given that Ann is saved from drowning. More formally, it seems plausible that $P(\mathring{\mathrm{A}}_i) = P(\mathrm{G}_i(V) \mid \mathrm{A})$. And that would mean that the equality in (12) must hold.

But now we might seem to be faced with another negative result due to David Lewis: his famous 'triviality result' against the so-called Adams's thesis (see e.g. Lewis 1976 and 1986, and Hájek and Hall 1994). Lewis's main target in his triviality argument was the idea that the probability of an indicative conditional, A → B, is identical to the conditional probability of B given A.[25] However, his result is easily generalised to a refutation of any claim of the form that for any probability function P and propositions A and B, there exists a proposition C such that $P(\mathrm{A} \mid \mathrm{B}) = P(\mathrm{C})$. So in particular it refutes the claim that there exists a proposition $\mathring{\mathrm{A}}_i$ such that $P(\mathring{\mathrm{A}}_i) = P(\mathrm{G}_i(V) \mid \mathrm{A})$, for any P, $\mathrm{G}_i(V)$ and A. For this reason, Broome 1991 insists that we must resist the temptation to identify the probability that A is good to some degree i with the conditional probability

[24] A generalised causal version of DAB could also be considered. Everything we say here about the evidential GDAB also holds for a causal GDAB, since, as we saw in §5, causal decision theory entails a causal DAE thesis.

[25] It is worth noting that Lewis accepted Adams's (1975) own view that the *assertability* of an indicative conditional equals the corresponding conditional probability (as a referee reminds us). What Lewis did not accept, however, was the view that the probability of an indicative's *truth* is equal to the corresponding conditional probability. Adams himself did not accept the latter view either, but the view has nevertheless come to be called 'Adams's thesis' (or sometimes 'Stalnaker's thesis', or just the 'Thesis').

that the world is good to degree i given A.

However, the aforementioned triviality results all depend on taking the contents of the propositions in question to be fixed. In particular, learning some proposition is not supposed to change the content of any other proposition (nor of that same one). But in §5 we argued that, first, DAE entails that this is precisely what happens with the goodness-level 'propositions' when one changes ones mind about how good some proposition is, and, second, that this is what must happen if DAE is not to imply the false Invariance principle.

The contents of the $\mathring{\mathrm{A}}_i$ propositions must not be fixed either, for similar reasons. These propositions also partition the space of possible worlds, but when a person learns that A is true, for instance, then unless A was already considered neither more nor less desirable than the tautology, A's place within the $\mathring{\mathrm{A}}_i$ partition must (if GDAB is true) shift such that its expected goodness becomes equal to that of the tautology (recall our discussion of Invariance from §3.2). But then since the content of A does not change when an agent learns that the proposition is true, the content of some $\mathring{\mathrm{A}}_i$ proposition must change when an agent learns this. So the contents of both $\mathring{\mathrm{A}}_i$ and $\mathrm{G}_i(V_{\mathrm{A}})$ change as an agent learns new information. The upshot is that the triviality arguments like those that have been been taken to undermine Adams's thesis do not invalidate the above argument for GDAB. For these arguments only work when the contents of the propositions involved are fixed.

A similar treatment can be given of another apparent problem for the generalised DAB thesis, namely that it seems that, given DAE, the value propositions $\mathring{\mathrm{A}}_i$ must be probabilistically independent of A, which then again entails the false Invariance principle. For since $\sum_i i.P(\mathrm{G}_i(V_{\mathrm{A}}) \mid \mathrm{A})$ might seem invariant under changes in the probability of A, it follows that if both DAE and GDAB are true, then $\sum_i P(\mathring{\mathrm{A}}_i)$ must also be invariant under changes in A. But then by GDAB, $V(\mathrm{A})$ must also be invariant under changes in the probability of A, which we have seen to be inconsistent with Jeffrey's understanding of desirability.

The above worry is mistaken however. In §5 we showed that $\sum_i i.P(\mathrm{G}_i(V_{\mathrm{A}}) \mid \mathrm{A})$ is *not* invariant under changes in the probability of A. It only seems to be so because of an implicit assumption that the contents of the goodness-level ‘propositions’ are fixed. But we have shown, first, that one should not make that assumption, and, second, that our formulation of DAE does not entail it. But then if both DAE and GDAB are true, $\sum_i P(\mathring{A}_i)$ may also change when an agent learns that A is true, just as the falsity of Invariance entails. In sum, the Desire-as-Expectation* thesis does not contradict the Generalised Desire-as-Belief thesis, given the indexical nature of the goodness-level propositions.

7. Concluding remarks

Contrary to what David Lewis thought, Bayesian decision theory does not rule out the possibility that there should be some particular quantitative relationship between a rational person’s desires and her evaluative beliefs. Although the simple version of the Desire-as-Belief thesis that he proposed is refuted by our counterexample, the more plausible Desire-as-Expectation* thesis—which should be acceptable to Humeans, anti-Humeans, subjectivists and objectivists—is not only consistent with Bayesian decision theory but entailed by it. And this thesis, in turn, is consistent with a version of DAB that allows for multiple goodness levels. Nor, contrary to appearances, do these theses imply the dubious Invariance principle since propositions expressing goodness claims must themselves have contents that vary. So Bayesian decision theory, it would seem, is general enough to allow for a range of different theories of value and human motivation.[26]

References

Adams, Ernest W. 1975, *The Logic of Conditionals* (Dordrecht: D. Reidel Publishing Company)

Bradley, Richard 1999, ‘Conditional Desirability’, *Theory and Decision*, 47, pp. 23–55

Bradley, Richard and Christian List 2009, ‘Desire-as-Belief Revisited’, *Analysis*, 69, pp. 31–7

[26] We would like to thank John Collins, Alan Hájek, Conrad Heilmann, Jim Joyce, Wlodek Rabinowicz, Katie Steele, Robbie Williams, three referees and an editor for *Mind* for helpful comments on different versions of this paper.

Broome, John 1991, ‘Desire, Belief and Expectation’, *Mind*, 100, pp. 265–7

Byrne, Alex and Alan Hájek 1997, ‘David Hume, David Lewis, and Decision Theory’, *Mind*, 106, pp. 411–28

Collins, John 1988, ‘Belief, Desire, and Revision’, *Mind*, 97, pp. 333–42

Costa, Horacio A., John Collins, and Isaac Levi 1995, ‘Desire-as-Belief Implies Opinionation or Indifference’, *Analysis*, 55, pp. 2–5

Daskal, Steven 2010, ‘Absolute Value as Belief’, *Philosophical Studies*, 148, pp. 221–9

Eells, E., B. Skyrms and E. W. Adams (eds) 1994, *Probability and Conditionals: Belief Revision and Rational Decision* (Cambridge: Cambridge University Press)

Gibbard, Alan and William L. Harper 1981, ‘Counterfactuals and Two Kinds of Expected Utility Theory‘, In Harper, Stalnaker, and Pearce 1981, pp. 153–90

Hájek, Alan and Ned Hall 1994, ‘The Hypothesis of the Conditional Construal of Conditional Probability,’ in Eells, Skyrms and Adams 1994, pp. 75–112

Hájek, A. and P. Pettit 2004, ‘Desire Beyond Belief’, *Australasian Journal of Philosophy*, 82, pp. 77–92

Harper, W. L., R. Stalnaker, and G. Pearce (eds) 1981, *Ifs: Conditionals, Belief, Decision, Chance, and Time* (Dordrecht: D. Reidel Publishing Company)

Jeffrey, Richard 1965, *The Logic of Decision*, reprinted 1990 (Chicago: The University of Chicago Press)

Jeffrey, Richard 1977, ‘A Note on the Kinematics of Preference’, *Erkenntnis*, 11, pp. 135-141

Joyce, James M. 1999, *The Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press)

Lewis, David 1976, ‘Probabilities of Conditionals and Conditional Probabilities’, *Philosophical Review*, 85, pp. 297–315

——1981, ‘Causal Decision Theory’, *Australasian Journal of Philosophy*, 59, pp. 5–30

——1986, ‘Probabilities of Conditionals and Conditional Probabilities II’, *Philosophical Review*, 95, pp. 581–589

——1988, ‘Desire as Belief’, *Mind*, 97, pp. 323–32

——1996, ‘Desire as Belief II’, *Mind*, 105, pp. 303–13

Nissan-Rozen, Ittay forthcoming, ‘A Triviality Result for the “Desire by Necessity” Thesis’, forthcoming in *Synthese*

Nozick, Robert 1969, ‘Newcomb’s Problem and two Principles of Choice’, in Rescher 1969, pp. 114–146

Oddie, Graham 1994, ‘Harmony, Purity, Truth’, *Mind*, 103, pp. 451–472

——2001, ‘Hume, the BAD Paradox, and Value Realism’, *Philo*, 4, pp. 109–22

Plato 360/1953 BC, *Symposium* trans. Benjamin Jowett (Cambridge: Pearson)

Rescher, N., (ed) 1969, *Essays in Honor of Carl G. Hempel* (Dordrecht: D. Reidel Publishing Company)

Russell, Bertrand 1905, ‘On Denoting’, *Mind*, 14, 479–93

Smith, Michael 1987, ‘The Humean Theory of Motivation’, *Mind*, 96, pp. 36–61

——1994, *The Moral Problem* (Oxford: Wiley-Blackwell)

Stefánsson, H. Orri 2014, ‘Desires, Beliefs and Conditional Desirability’, *Synthese*, 191, pp. 4019–35

Weintraub, Ruth 2007, ‘Desire as Belief, Lewis Notwithstanding’, *Analysis*, 67, pp. 116–22

Williams, J. Robert G. 2010, ‘Counterfactual Desire as Belief’, unpublished manuscript