Intrinsic Desire as Belief

Wolfgang Schwarz

Draft, May 2025

1 Introduction

We have beliefs about what is good or bad, about what reasons we have, about what we should do. These normative beliefs are connected to desire and motivation. Most people desire what they believe is good, and not what they believe is bad; they are motivated to make the world better, not worse. This connection seems to contradict Humean ideas about the inertness of belief. [Lewis 1988, 1996] argued that it even clashes with elementary principles of decision theory.

Lewis's argument is widely thought to rest on an obvious mistake – although there is no agreement on what that mistake is. Among other things, Lewis has been accused of misconstruing the form of the belief-desire link (e.g., [Price 1989], [Broome 1991]), relying on a faulty "invariance" assumption for the dynamics of desire (e.g., [Bradley and List 2009], [Bradley and Stefánsson 2016]), overlooking the "absoluteness" ([Weintraub 2007], [Daskal 2010]) or "indexicality" ([Hájek and Pettit 2004]) of normative judgements, and relying on a misguided "evidential" conception of motivation and desire (e.g., [Byrne and Hájek 1997], [Oddie 2001], [Collins 2015]). I'll argue that while some of these objections do point at shortcomings in Lewis's original argument, they don't help escape the underlying problem. There really are serious obstacles to formalizing the connection between normative belief and desire.

Nonetheless, it is possible. To do so, we need to focus on *intrinsic* desire and beliefs about *intrinsic* goodness. I'll show that this not only avoids the Lewisian problems; it also diffuses the threat of "fetishism", raised in [Smith 1994]: it explains why an agent who cares about normative matters needn't have a single intrinsic desire, to bring about what is good.

2 Desire As Belief

Let's briefly review Lewis's argument. We want to connect normative beliefs to desire. Following Lewis, let's assume that the relevant beliefs are concerned with what is or isn't "good". We can treat this as a placeholder: 'A is good' means that A has whatever positive normative status we want to connect to desire. We assume that goodness applies to (possibly centred) propositions, which may describe actions, intentions, events, or states of affairs.

Belief and desire come in degrees. In the Bayesian tradition of [Jeffrey 1965], they are represented

by a credence function Cr and a desirability function V. Rational credence is assumed to satisfy the rules of probability; desirability is assumed to satisfy an averaging principle, on which more soon.

Goodness may also come in degrees. We therefore need to connect $V(A)$, an agent's degree of desire for A, to the agent's graded beliefs about A's degree of goodness. A natural way to do this identifies $V(A)$ with the credence-weighted average (in probability jargon, the "expectation") of A's goodness. That is, if $g(A) = x$ is the proposition that A is good to degree x, the suggestion would be that for all rational credence functions Cr, desirability functions V, and propositions A,[1]

$$V(A) = E[g(A)] = \sum_x x \ \mathrm{Cr}(g(A) = x). \tag{DAE}$$

("DAE" stands for "Desire As Expectation".)

To illustrate, consider the problem of decision-making under moral uncertainty. Suppose you are undecided between two moral theories that disagree about the value of some proposition A – say, that you donate 10% of your income to charity. Theory One says that you should do it, Theory Two regards it as optional. We might expect that the more strongly you believe in the first theory, the more you are motivated to donate. More precisely, if each theory assigns a goodness score to A, we might expect that your desire for A is proportional to the average of the two scores, weighted by your degree of belief in each theories. This is a popular idea in the moral uncertainty literature (see, e.g., [Lockhart 2000], [MacAskill et al. 2020a]). It appears to be an instance of DAE.

For another illustration, suppose you are somewhat confident that you have a strong reason to bring about A, but you can't rule out the possibility that you have no such reason. Again, we might expect that the more confident you are that you have a reason, and the stronger that believed reason, the greater your total degree of motivation towards A. [Gregory 2021: ch.8], for example, suggests that (at least under ideal conditions), your total degree of motivation towards A should be the credence-weighted average of the strength of your reasons for A, in line with DEA.[2]

So DAE is a natural conjecture. It is not a strawman. Yet it appears to run into serious trouble. To keep the maths simple, Lewis mostly concentrates on the case where there are only two degrees of goodness: 0 and 1, say. DAE then reduces to $V(A) = \mathrm{Cr}(g(A) = 1)$. Writing $\mathring{A}$ for $g(A) = 1$, we get the "Desire As Belief" thesis that is Lewis's main target:

$$V(A) = \mathrm{Cr}(\mathring{A}). \tag{DAB}$$

This *is* a bit of a strawman, as it's hard to comprehend how someone could distinguish only two levels of goodness. Imagine, following [Bradley and Stefánsson 2016], that Alice and Bob are in danger of drowning. One might think that rescuing both is better than rescuing only one, which is

1 In what follows, displayed equations are always assumed to be closed by universal quantification over Cr, V and the propositions A, B, etc. If an equation involves a conditional credence $\mathrm{Cr}(A/B)$, it is assumed to be restricted to cases where $\mathrm{Cr}(B) > 0$. I assume that Cr is a discrete distribution over $g(A)$. We could allow for continuous distributions by using densities and integrals. I'll stick to the discrete case for simplicity.

2 Gregory suggests that if you give positive credence to having a reason of strength x_1 and having a reason of strength x_2, then you have two desires towards A, one with "strength" x_1, the other with "strength" x_2. Your degree of credence in the reason hypotheses doesn't affect the "strength" of the desires, but it affects the net motivational force of the combined desires, which (according to the somewhat tentative proposal in [Gregory 2021: 159]) is given by DAE.

still better than rescuing neither. But this is impossible if there are only two levels of goodness. If rescuing one is better than rescuing neither, then rescuing both can't be better than rescuing one. Bradley and Steffánson present this as a "counterexample" to DAB, assuming that one may rationally prefer rescuing both to rescuing one (and rescuing one to rescuing neither). But Lewis assumes (again for simplicity) that the agent only cares about goodness. If you really believe that rescuing both is no better than rescuing one, and you only care about goodness, why would you prefer to rescue both? The case isn't a counterexample to DAB, but it illustrates how bizarre it would be to only distinguish two levels of goodness. That said, the bizarre assumption really does simplify the formalities – which is why I'll mostly stick to it in what follows.

Here, then, is Lewis's refutation of DAB, as presented in [Lewis 1996]. Take any instance of DAB. Now suppose the agent whose attitudes are represented by Cr and V receives the information that A is true. Let Cr^A be the new credence function and V^A the new desire function. Given that Cr^A comes from Cr by conditionalizing on A, we have

$$\mathrm{Cr}^A(\mathring{A}) = \mathrm{Cr}(\mathring{A} \mid A). \qquad \text{(Conditionalization)}$$

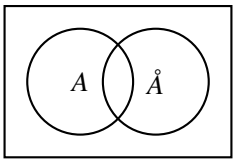
Assuming that the desirability of a proposition does not change by learning that it is true, we also have

$$V^A(A) = V(A). \qquad \text{(Invariance)}$$

DAB requires that $V^A(A) = \mathrm{Cr}^A(\mathring{A})$. So $\mathrm{Cr}(\mathring{A}) = V(A) = V^A(A) = \mathrm{Cr}(\mathring{A} \mid A)$. For short,

$$\mathrm{Cr}(\mathring{A}) = \mathrm{Cr}(\mathring{A}/A) \qquad \text{(Independence)}$$

So DAB implies that $\mathring{A}$ is (probabilistically) independent of A.



But picture A and $\mathring{A}$ as regions in logical space. One would think that A and $\mathring{A}$ can generally be true together. So the regions overlap. $\mathrm{Cr}(\mathring{A}/A)$ is the ratio of the credence in their intersection over the credence in A (at least if $\mathrm{Cr}(A) > 0$). Independence says that this ratio equals the total credence in $\mathring{A}$. That might be true for a particular credence function Cr, but if it is, we can easily make it false. For example, if we move credence from outside $A \vee \mathring{A}$ into $\mathring{A} \wedge \neg A$, or if we conditionalize on $A \vee \mathring{A}$, we increase $\mathrm{Cr}(\mathring{A})$ without changing $\mathrm{Cr}(\mathring{A}/A)$. Conversely, we can keep $\mathrm{Cr}(\mathring{A})$ constant while changing $\mathrm{Cr}(\mathring{A}/A)$ by, say, moving credence from outside $A \vee \mathring{A}$ into $A \wedge \neg\mathring{A}$. That is, even if Independence holds for a particular credence function, it will fail for others. So DAB – as a hypothesis about *all* rational credence functions – is false.

In fact, the argument doesn't just refute an unrestricted version of DAB. It shows that the equality $V(A) = \mathrm{Cr}(\mathring{A})$ can't even hold for a single proposition A, provided that it is logically compatible with $\mathring{A}$, and for a very narrow set of Cr, V pairs that is closed under conditionalizing on A, satisfies Invariance, and allows for some variation in how Cr is distributed over the possible combinations of

A and $\mathring{A}$. It wouldn't help, for example, to insist that not any old proposition can be regarded as good, or that the link between normative belief and desire only holds under "ideal conditions", provided the conditions are closed in the relevant way. (As [Lewis 1988] and [Byrne and Hájek 1997] show, one can substantially weaken the closure conditions on Cr.)

Note also that very few assumptions from decision theory are involved. We didn't assume, for example, that rational agents always maximize expected utility. We assumed that degrees of belief and desire can be represented by a credence function Cr and a desirability function V, but this much is already presupposed by the formulation of DAB. It's hard to see how moving to imprecise credence and desire could help. Indeed, [Collins 1988] shows (albeit with a rather different style of argument) that even an ungraded formulation of DAB runs into trouble.

What if we drop the simplifying assumption that there are only two levels of goodness? [Lewis 1988] explains how the refutation generalizes to DAE.[3] Briefly, if DAE holds before and after conditionalizing on A, and Invariance holds, we get an equation between an unconditional and a conditional expectation:

$$\sum_x \mathrm{Cr}(g(A) = x)\ x = \sum_x \mathrm{Cr}(g(A) = x \ / \ A)\ x.$$

Like Independence, this may hold for a particular credence function, but it is easy to break by moving around credence.

If we want to escape the refutation, we have to either give up Invariance, or the closure conditions on Cr, or revise the link between belief and desire. Unfortunately, Lewis's argument sheds little light on *why* DAB and DAE run into trouble. One is tempted to blindly tinker with the assumptions and hope that things will work out. Before we look at such tinkering, let's try to get a clearer understanding of where the trouble comes from.

3 Averaging

Think again of propositions as regions in logical space. Most propositions have more desirable and less desirable parts. Within the region in which you win the lottery, for example, we find subregions where you lead a happy and fulfilling life, and others in which you end up miserable and alone. Similarly for the regions in which you donate to charity, rescue a drowning child, or consult a doctor about a medical problem.

How desirable is such an "uneven" proposition, with more desirable and less desirable parts? It depends on the probability of the parts. If you desire to win the lottery, you probably assign greater probability to the subregions in which you lead a happy life than to the subregions in which you're miserable and alone. This motivates the averaging principle in Jeffrey's account of desire. The principle says that if A divides (exhaustively and exclusively) into A_1 and A_2, then A's desirability is the average of the desirability of these parts, weighted by their probability conditional on A:

$$V(A) = V(A_1)\,\mathrm{Cr}(A_1/A) + V(A_2)\,\mathrm{Cr}(A_2/A). \qquad \text{(Averaging)}$$

3 Lewis also explains how his result generalizes to principles according to which beliefs about goodness are only *a* source of motivation, besides others.

We use the probability of the parts conditional on A, rather than their unconditional probability, so that the weights add up to 1: we don't want to say that the desirability of A is near zero merely because A is improbable.

Let's pretend that the algebra over which Cr is defined is finite. Averaging then entails that the desirability of a region A is the credence-weighted average of the desirability of the "worlds" in the region, where a world is a maximally specific proposition (an atom in the algebra):

$$V(A) = \sum_{w} V(w)Cr(w/A).$$

The V function restricted to worlds, sometimes written V_b, represents the agent's *basic desires*. Since worlds don't have subregions, basic desires aren't sensitive to how credence is distributed over subregions.

Next, we need to know how basic desires relate to beliefs about goodness. We assume that Cr is defined, among other things, over normative propositions. So a "world" is a maximally specific proposition about both descriptive and normative matters.

I assume that a maximally specific normative proposition can be represented as a function v that assigns a goodness level to each world.[4] Let v_w be the value function built into world w. If your only basic desire is for goodness, you should arguably desire each world in proportion to the goodness level it assigns to itself:

$$V_b(w) = v_w(w). \qquad \text{(Fetishism)}$$

Consider, for example, a world in which (a) the only ultimate good is that the number of stars is even, and (b) the number of stars is indeed even. Plausibly, this is a world where your desire is satisfied: it's a world where everything is as it ought to be, and that's what you want. (Compare [Lewis 1988: p.332].)

The label 'fetishism' is a nod to [Smith 1994: 60-91], who argues that an agent whose only basic desire is for goodness would have a "moral fetish". I'll return to this point in sections 6 and 9, where I'll also describe a non-fetishist way in which one might care about goodness. But Fetishism is the simplest, most straightforward model, and it will be useful to see where it leads.

If there are only two levels of goodness, each world is either good according to itself or not good according to itself. Let G be the set of worlds that are good according to themselves. Informally, G says that the world is good, or that everything is as it ought to be. Fetishism then simplifies to[5]

$$V_b(w) = \begin{cases} 1 & \text{if } w \in G \\ 0 & \text{if } w \notin G. \end{cases}$$

(As Lewis points out in [Lewis 1996: sec.6], this follows from DAB together with the *Stability* assumption that conditionalization doesn't affect basic desires. For let Cr^w, V^w result from Cr, V by

4 This allows for norms that don't just pertain to descriptive matters but also to normative matters: according to some norms, it may be good that donating to charity is good. It also allows for norms that only pertain to descriptive matters. In that case, the goodness of each world (relative to these norms) is determined by its descriptive aspects.

5 Fetishism, especially in this simplified formulation, is closely related to a principle that [Lewis 1996] calls 'Desire by Necessity'. See section 7.

conditionalizing on w. By Stability, $V^w(w) = V(w)$. By DAB, $V^w(w) = \mathrm{Cr}^w(\mathring{w}) = \mathrm{Cr}(\mathring{w}/w)$. If $w \in G$ then $w \models \mathring{w}$, otherwise $w \models \neg\mathring{w}$. So $\mathrm{Cr}(\mathring{w}/w) = 1$ if $w \in G$ else $\mathrm{Cr}(\mathring{w}/w) = 0$. And so $V(w) = 1$ if $w \in G$ else $V(w) = 0$.)

Let's assume Fetishism. Combined with Averaging, we get

$$V(A) = Cr(G/A).$$

Together with the DAB equation $V(A) = \mathrm{Cr}(\mathring{A})$, this yields an equality of belief with conditional belief:

$$\mathrm{Cr}(\mathring{A}) = \mathrm{Cr}(G/A). \tag{BACB}$$

We'll see that BACB threatens to clash with standard probability theory. But on the face of it, it looks plausible. If you think that donating to charity is good (that you have a reason to do it, etc.), you probably assign greater credence to the good subregions of that proposition than to the bad ones. If your credence gradually shifts towards the bad subregions, you'll become increasingly less confident that donating to charity is good.[6]

Of course, it is unintuitive to have just two levels of goodness. With more than two levels of goodness, Fetishism and Averaging yield a version of the "Desire As Expected Goodness" thesis defended in [Broome 1991]:

$$V(A) = \sum_x xCr(G_x/A).$$

If we combined this with DAE (the generalized form of DAB for more than two levels of goodness), we get another equation between an unconditional and a conditional expectation:

$$\sum_x \mathrm{Cr}(g(A) = x)\, x = \sum_x \mathrm{Cr}(G_x/A)\, x. \tag{EACE}$$

This leads to the same kind of trouble as BACB.

4 Triviality

Return to the simple case with two levels of goodness. Like Independence, BACB identifies a conditional probability with an unconditional probability.

$$\mathrm{Cr}(\mathring{A}) = \mathrm{Cr}(G/A). \tag{BACB}$$

As before, this equality may hold for a particular credence function, but it can't hold in general. Suppose you give a small amount of credence to A, which has a good part and a bad part. Within A, most of your credence lies on the good part. By BACB, you must be fairly confident that A is

6 Weintraub [2007] gives a supposed counterexample. There are two lotteries, each with a prize of $1000, but the chance of winning is greater in one than in the other. Weintraub says that winning the easy lottery is as good as winning the hard lottery, and that not winning either is as good as not winning the other, even though the probabilities of winning are different. Averaging, however, implies that if $V(A) = V(B)$ and $V(\neg A) = V(\neg B)$ then $\mathrm{Cr}(A) = \mathrm{Cr}(B)$. Likewise, BACB implies that if you're sure that A and B are equally good, as are their negations, then you must deem them equally likely. This may initially seem odd, but I think it makes sense. In the lottery case, it's better to win the hard lottery:

good. If we move around your credence within A, so that most of it comes to lie on the bad part, you must become confident that A is not good. We've only moved a small fraction of your credence, the fraction that lies in A. Yet most of your overall credence now lies outside $\mathring{A}$, while previously most of it lay inside. This is impossible!

Unlike Lewis's argument, this argument doesn't involve any updating. I'm not comparing an earlier and a later credence function. I haven't assumed Invariance. The problem arises because we seem to evaluate whether an uneven proposition is good by looking at the comparative probability of its good and bad parts. To evaluate whether it's good if you win the lottery, we look at what is likely to happen if you win it. If we're confident that you'll be happy, we think winning is good; if we're confident that you'll be miserable, winning is bad. This looks innocent, but it can't be true!

The underlying problem doesn't rely on DAB. It generalizes to DAE[7] any beyond. Consider a comparative version of DAB:

$$V_1(A) > V_2(A) \text{ iff } \mathrm{Cr}_1(\mathring{A}) > \mathrm{Cr}_2(\mathring{A}). \qquad \text{(Comparative DAB)}$$

Here, V_1, Cr_1 and V_2, Cr_2 are two pairs of a credence function and a desirability function. By Averaging and Fetishism, the Left-to-Right direction of Comparative DAB entails that

$$\text{If } Cr_1(G/A) > Cr_2(G/\neg A) \text{ then } Cr_1(\mathring{A}) > Cr_2(\mathring{A}).$$

This looks plausible (modulo the simplifying assumption that there are only two levels of goodness). Suppose you and I disagree on what happens if you win the lottery: I am confident that good things will happen, you are confident that bad things will happen. Surely I will be more confident that winning is good.

But assume Cr_1 and Cr_2 both give low credence to A, and $\mathrm{Cr}_1(G/A) > \mathrm{Cr}_2(G/\neg A)$. If $\mathring{A}$ extends beyond A (as it must if $\mathrm{Cr}_1(\mathring{A})$ can be high), we can make $\mathrm{Cr}_2(\mathring{A})$ exceed $\mathrm{Cr}_1(\mathring{A})$ by focusing Cr_2 outside A on $\mathring{A}$ and Cr_1 outside A on $\neg\mathring{A}$. (If $\mathring{A}$ doesn't extend beyond A, we can make Cr_1 arbitrarily small by reducing $\mathrm{Cr}_1(A)$ while preserving $\mathrm{Cr}_1(G/A)$; unless $\mathrm{Cr}_2(\mathring{A})$ is zero, we'll reach a point where $\mathrm{Cr}_1(\mathring{A}) < \mathrm{Cr}_2(\mathring{A})$.) So Comparative DAB can't hold in general.

We can also consider direct comparative judgements. Let $A \succ B$ be the hypothesis that A is better than B. One might have thought that if you're sure that A is better than B, and you only care about goodness, then you'll prefer A to B:[8]

$$\text{If } \mathrm{Cr}(A \succ B) = 1 \text{ then } V(A) > V(B). \qquad \text{(Better)}$$

this comes with a greater probability of winning both. Imagine an extreme case where the chance of winning the easy lottery is 0.99, while the chance of winning the hard lottery is 0.01. Winning the hard lottery is then equivalent to the prospect of getting \$2000 with probability 0.99 and \$1000 with probability 0.01; winning the easy lottery amounts to getting \$2000 with probability 0.01 and \$1000 with probability 0.99. Surely the first is better. (If we compare winning *only* the easy lottery and winning *only* the hard lottery, their negations aren't equally good: not winning the easy lottery is worse, because it implies a greater probability of winning neither lottery.)

7 The same problem affects EACE, where we have more than two levels of goodness: $\sum_x \mathrm{Cr}(G_x/A)\, x$ only depends on the relative distribution of credence inside A; if $\mathrm{Cr}(A) < 1$, we can vary the distribution outside A to break the equality with $\sum_x \mathrm{Cr}(g(A) = x)\, x$.

8 [Russell and Hawthorne 2016: 315f.] consider a closely related principle ("Comparative Value"). They use a dynamic

By Averaging and Fetishism, we could infer that

$$\text{If } Cr(A \succ B) = 1 \text{ then } Cr(G/A) > Cr(G/B).$$

This, too, looks plausible: if you're sure that A is better than B you must think it's more likely that good things happen given A than given B. Alas, it can't hold in general. We need to distinguish two cases. If $A \succ B$ overlaps G, let Cr be concentrated on $(A \succ B) \wedge G$; we'll have $\mathrm{Cr}(A \succ B) = 1$, but $\mathrm{Cr}(G/A) = \mathrm{Cr}(G/B) = 1$. If $A \succ B$ does not overlap G, $\mathrm{Cr}(A \succ B) = 1$ entails $\mathrm{Cr}(G/A) = \mathrm{Cr}(G/B) = 0$.

These are "triviality arguments" because they show that the suggested principles can only hold under conditions that would make them trivial. For example, DAB, Comparative DAB, and Better are trivially satisfied if G is the set of all worlds, so that all propositions are good.

I want to emphasize how perplexing these arguments are. They don't just refute a dubious combination of contentious hypotheses from metaethics and decision theory. If successful, they refute assumptions that seem obviously true.

The grandfather of all triviality arguments is Lewis's [1976] refutation of *Stalnaker's Thesis*, the conjecture that the probability of an indicative conditional 'if A, B' (in symbols, $A \Rightarrow B$) equals the conditional probability of B given A:

$$\mathrm{Cr}(A \Rightarrow B) = \mathrm{Cr}(B/A). \qquad \text{(Stalnaker's Thesis)}$$

This, too, has seemed obviously true to many. To evaluate the probability of $A \Rightarrow B$, we seem to evaluate the probability of B on the supposition that A. But it fails for the same reason as BACB, which is a special case of Stalnaker's Thesis: it would mean that the probability of $A \Rightarrow B$ is determined by the relative probability of B within A, even if A is improbable.

Similar triviality arguments have been constructed for 'must' and 'might', on their epistemic reading (see, e.g., [Russell and Hawthorne 2016], [Goldstein and Santorio 2021]). All these cases – conditionals, epistemic modals, statements about goodness – have something in common. Intuitively, the relevant statements make a claim about the available information rather than directly about the world. 'Might A' seems to say that the available information is compatible with A; 'if A, B' seems to say (roughly) that all A-worlds compatible with the available information are B-worlds; 'A is good' seems to have a reading on which it says (roughly) that all A-worlds compatible with the available information are good. In semantics, statements of this kind are often called *information-sensitive*. Unlike ordinary statements, their evaluation seems to require an information parameter rather than a world parameter. Oddly, when we assess the probability of an information-sensitive statement, we tend to fill the information parameter with the very probability measure that we use for the assessment. (The linguistic environment can further manipulate that parameter; we'll see examples in the next section.)

The exact mechanisms of why this happens are still a matter of debate. As far as I can tell, there is, however, a near consensus that one can't give an adequate account of information-sensitive ex-

argument to show that it leads to $\mathrm{Cr}(A \wedge B \wedge (A \succ B)) = 0$, which they find is problematic. I agree. But the consequence is true for a wide range of credence functions. So theirs is not quite a triviality result.

pressions within the confines of standard truth-conditional semantics: one can't explain our use of information-sensitive statements by assuming that they simply pick out a set of worlds. (See, among many others, [Yalcin 2011], [Moss 2015], [Stalnaker 2019], [Hawke and Steinert-Threlkeld 2021], [Mandelkern 2024], [Ciardelli and Ommundsen 2024], [Rudin 2025].)[9]

This raises a problem. If statements about goodness don't express propositions, we can hardly find a link between beliefs towards these (non-existent) propositions and desire. We may find true *sentences* that appear to express such a link – sentences like DAB or Comparative DAB or Better –, but these sentences won't mean what they seem to mean. They won't express a genuine link between beliefs about goodness and desire.

5 Escape from triviality?

Let's revisit the argument against BACB, to see how it might be resisted. We've assumed that A divides into a G part and a $\neg G$ part, and that the overall probability of A is low. According to BACB, the probability of $\mathring{A}$ equals the fraction of the probability of A within G. If most of the probability in A lies on the $\neg G$ part, this means that the probability of $\mathring{A}$ is low. If we move the probability in A onto the G part, the probability of $\mathring{A}$ becomes high. But moving a small amount of probability inside A can't change any unconditional probability from low to high.

One way to resist this argument is to simply forbid any movement of credence that would render BACB false: we are only allowed to move around credence inside A if we also move around a lot of credence outside A. Another, more subtle way to resist the argument drops the assumption that $\mathring{A}$ is a fixed proposition. Instead, $\mathring{A}$ is a proposition with low probability if most of the probability in A lies on the $\neg G$ part; if we move that probability onto G, $\mathring{A}$ becomes a different proposition, with high probability.

Both of these ideas have been explored in the literature on conditionals, in response to the triviality arguments against Stalnaker's Thesis. (See, e.g., [Bradley 2012] and [Goldstein and Santorio 2021] for the first; [Kaufmann 2009] and [Bacon 2015] for the second; both suggestions go back to [van Fraassen 1976].) They face numerous problems.[10] Here, I only want to mention one kind of problem that they both share: by validating Stalnaker's Thesis, they make false predictions about conditional probabilities and probabilities of complex statements containing conditionals.

Let's look at an example from [Goldstein and Santorio 2021] and [Ciardelli and Ommundsen 2024], drawing on [Fitelson 2015]. A fair die has been tossed. Let *Low* mean that the outcome is below 4, *Even* that it is even, and *Two* that it is 2. What's the probability of $\mathit{Low} \Rightarrow \mathit{Two}$, on the assumption that the outcome is even? Intuitively, it's 1: $\mathrm{Cr}(\mathit{Low} \Rightarrow \mathit{Two}/\mathit{Even}) = 1$. By probability theory,

9 Early versions of this non-propositional view often adopted an "expressivist" approach: utterance of a conditional, for example, were suggested not to assert a proposition, but to express a high conditional credence in the consequent given the antecedent (e.g., [Adams 1975], [Edgington 1995], [Bennett 2003]).

10 With respect to goodness, [Hájek and Pettit 2004] argue that there are independent reasons to assume that 'good' (or 'right' etc.) is context-sensitive. This may be true, but it isn't enough to escape the trouble. We would need a very precise form of context-sensitivity by which $\mathring{A}$ always picks out some proposition whose probability, according to the speaker, matches that of G given A. One of the many further problems for this kind of view is that it arguably couldn't fit the standard model of context-sensitivity of [Kaplan 1989], according to which context-sensitive expressions have a non-context-sensitive "diagonal": the triviality arguments would re-appear for the diagonal.

this implies that $\mathrm{Cr}(\mathit{Low} \Rightarrow \mathit{Two}) \geq \mathrm{Cr}(\mathit{Even})$. Yet by Stalnaker's Thesis, $\mathrm{Cr}(\mathit{Low} \Rightarrow \mathit{Two}) = \mathrm{Cr}(\mathit{Two}/\mathit{Low}) = 1/3$, and obviously $\mathrm{Cr}(\mathit{Even}) = 1/2$. The intuitive judgement about the probability of $\mathit{Low} \Rightarrow \mathit{Two}$ given *Even* is probabilistically inconsistent with Stalnaker's Thesis.

Worse, the judgement about conditional probability does not stand alone. Consider the probability of $\mathit{Even} \Rightarrow (\mathit{Low} \Rightarrow \mathit{Two})$. This, too, is intuitively 1. By Stalnaker's Thesis, it follows that $\mathrm{Cr}(\mathit{Low} \Rightarrow \mathit{Two}/\mathit{Even}) = 1$, which we've seen is incompatible with Stalnaker's Thesis.[11] Or consider the probability of $\mathit{Even} \wedge (\mathit{Low} \Rightarrow \mathit{Two})$. Intuitively, this equals the probability of *Even*: $\mathit{Low} \Rightarrow \mathit{Two}$ doesn't seem to add anything to *Even*; if the outcome is even then it must be two if it is below four. But if $\mathrm{Cr}(\mathit{Even} \wedge (\mathit{Low} \Rightarrow \mathit{Two})) = \mathrm{Cr}(\mathit{Even})$ then, again, $\mathrm{Cr}(\mathit{Low} \Rightarrow \mathit{Two}/\mathit{Even}) = 1$, by the Ratio Formula for conditional probability. All these judgements are probabilistically incompatible with Stalnaker's Thesis.

In the same way, it is easy to think of cases where $\mathrm{Cr}(A \Rightarrow B/\neg A \vee B)$ is intuitively high, even though $\mathrm{Cr}(A/B)$ is considerably lower than $\mathrm{Cr}(A \vee B)$, or of cases where $\mathrm{Cr}(A \Rightarrow B/C)$ seems high and $\mathrm{Cr}(A \Rightarrow B/D)$ low, even though the probability of $B \vee D$ is moderately high.[12] These judgements, too, are probabilistically inconsistent with Stalnaker's Thesis.

Unlike the standard triviality arguments, these arguments don't apply Stalnaker's Thesis after moving around credence. They don't involve any movement of credence at all. As such, neither the ban on moving credence nor the "changing meaning" response can escape them.

We can construct similar static arguments against BACB and other DAB-type principles. I'll give three.

First. Let A be the pessimistic assumption that greenhouse gases will keep rising for the next 50 years. Let B be the more optimistic assumption that *either* greenhouse gases will stop rising within 50 years *or* greenhouse gases have a net positive effect on the environment. Suppose we think it's likely that greenhouse gases will stop rising, but highly unlikely that they are net positive. So $\mathrm{Cr}(B)$ is high. By BACB, it follows that $\mathrm{Cr}(\mathring{A})$ is low. (This is intuitively correct, on an information-sensitive reading of 'good': it's not good if greenhouse gases keep rising.) On the other hand, $\mathrm{Cr}(\mathring{A}/B)$ is intuitively high: conditional on B, runaway greenhouse emissions are sure to have good consequences. But these judgements are probabilistically inconsistent, as probability theory requires that $\mathrm{Cr}(\mathring{A}/B) \leq \mathrm{Cr}(\mathring{A}) / \mathrm{Cr}(B)$.[13]

Second. A and B are as before. Intuitively, we could well be sure that B is better than A: $\mathrm{Cr}(B \succ A) = 1$. Now consider the conjunction of A and B. This expresses the unlikely hypothesis that greenhouse gases are net positive. The hypothesis is unlikely, but not completely impossible: $\mathrm{Cr}(A \wedge$

11 One might even suggest a general ("Import/Export") principle: $\mathrm{Cr}(A \Rightarrow (B \Rightarrow C) = \mathrm{Cr}(C/A \wedge B)$. More cautiously, one might suggest that $\mathrm{Cr}(A \Rightarrow (B \Rightarrow (A \wedge B))) = 1$, or that $\mathrm{Cr}(B \Rightarrow (A \wedge B)/A) = 1$. Intuitively, these are at least as plausible as Stalnaker's Thesis.

12 The first kind of case is probabilistic version of the notorious Or-to-If inference, discussed e.g. in [Stalnaker 1975]. A standard instance of the second kind of case is the riverboat case from [Gibbard 1981].

13 The counterexample loses some of its intuitive force by the implausible commitment to only two levels of goodness. As usual, though, we can drop this assumption. In the example, $E[g(A)]$ is intuitively low, $\mathrm{Cr}(B)$ is high, and $E[g(A)/B]$ is high. But this is impossible. To see why, assume that $g(A)$ is bounded in [0,1] and consider the mathematically easy (and not implausible) case where $E[g(A)/B] = 1$. By the law of total expectation, $E[g(A)] = E[g(A)/B] \cdot \mathrm{Cr}(B) + E[g(A)/\neg B] \cdot \mathrm{Cr}(\neg B)$. Given $E[g(A)/B] = 1$, it follows that $E[g(A)] = \mathrm{Cr}(B) + E[g(A)/\neg B] \cdot \mathrm{Cr}(\neg B)$. But if $g(A)$ is bounded in [0,1], $E[g(A)/\neg B] \geq 0$. So $E[g(A)] \geq \mathrm{Cr}(B)$. This contradicts the assumption that $E[g(A)]$ is low and $\mathrm{Cr}(B)$ high.

$B) > 0$. Conditional on the unlikely hypothesis, B clearly isn't better than A: $\mathrm{Cr}(B \succ A/A \wedge B) = 0$. These three judgements are probabilistically inconsistent.[14]

Third, a Bayesian version of the miners puzzle (from [Kolodny and MacFarlane 2010]). Ten miners are trapped in a mine and threatened by rising water; we can block at most one of two shafts from being flooded. Without further information about the location of the miners, blocking neither shaft seems best: it has the greatest expected goodness. Conditional on the miners being in shaft A, however, blocking shaft A is best; conditional on the miners being in shaft B, blocking shaft B is best. These judgements about what is "best" (or better than the alternatives) contradict probability theory: if 'N is best' has high unconditional probability, it can't have low conditional probability conditional on each of two hypotheses whose disjunction has high probability.

These cases illustrate a phenomenon I've mentioned at the end of the previous section: when we assess the probability of an information-sensitive statement, we don't always fill the information parameter with that same probability measure; the linguistic environment can manipulate the parameter. In particular, when we assess the probability of such a statement on the supposition that so-and-so is the case, we include the supposition in the information parameter.[15]

One could, of course, try to work around this issue. We could revise probability theory. Or we could try to tell an error theory for the problematic judgements. But the prospects are beginning to look dim.

6 Invariance and Alignment

DAB identifies $\mathrm{Cr}(\mathring{A})$ with $V(A)$. This forces $\mathring{A}$ to be information-sensitive: by Averaging, $V(A)$ depends on the distribution of credence over the parts of A; to evaluate $\mathrm{Cr}(\mathring{A})$, we therefore need to know how Cr is distributed within A.

Averaging is part of Jeffrey's decision theory, which Lewis took for granted in his papers. But Lewis's refutation of DAB doesn't make use of Averaging. Nor did he appeal to Fetishism. The only assumption about V Lewis used is Invariance: that if Cr^A, V^A come from Cr, V by conditionalizing on A, then $V^A(A) = V(A)$.

Invariance is a consequence of Averaging and the assumption that conditionalizing on some information leaves an agent's basic desires V_b unchanged:

$$V_b^E(w) = V_b(w). \qquad \text{(Stability)}$$

For suppose we conditionalize the credence function Cr on some hypothesis E, without changing V_b.

14 Averaging entails that if $\mathrm{Cr}(A \wedge B) = 1$ then $V(A) = V(B)$. If judgements about betterness mirror comparative desire, it follows that $\mathrm{Cr}(A \succ B) = 0$ whenever $\mathrm{Cr}(A \wedge B) = 1$. In particular, $\mathrm{Cr}(A \succ \top) = 0$ whenever $\mathrm{Cr}(A) = 1$. [Weintraub 2007] objects that we don't become indifferent towards a proposition merely by learning that it is true. But indifference is measured by comparing a proposition to its negation, not by comparing it to the tautology. As $\mathrm{Cr}(A)$ approaches 1, $\mathrm{Cr}(A \succ \neg A)$ generally doesn't approach 0.

The argument in the text is related to the non-static argument against Better in [Russell and Hawthorne 2016: 315f.].

15 Compare the observation in [Yalcin 2007]: 'Suppose p and might not p' is infelicitous; here we evaluate 'might not p' relative to a body of information that includes the supposition p.

Then

$$\begin{aligned} V^E(A) &= \sum_w V_b^E(w)\,\mathrm{Cr}^E(w/A) && \text{[by Averaging]} \\ &= \sum_w V_b(w)\,\mathrm{Cr}^E(w/A) && \text{[by Stability]} \\ &= \sum_w V(w)\,\mathrm{Cr}(w/A \wedge E) && [\text{by } \mathrm{Cr}^E = \mathrm{Cr}(*/E)] \\ &= V(A \wedge E) && \text{[by Averaging].} \end{aligned}$$

For the special case where A = E, this implies that $V^A(A) = V(A \wedge A) = V(A)$.

Conversely, however, Invariance and Stability don't entail Averaging. For example, Invariance still holds if we replace Averaging with "risk-weighted" forms of averaging that give more weight to the especially bad parts of a proposition. This still makes V information-sensitive; it still leads to Triviality.

The source of the triviality problem was that $V(A)$ only depends on the credence conditional on A. This follows from Averaging, but it is already entailed by Invariance: if $V(A)$ equals $V^A(A)$, it can't depend on how credence is distributed outside A, as that part of Cr leaves no trace after conditioning on A. (Technically, Invariance allows that $V(A)$ doesn't depend on the distribution of credence within A either, but then it wouldn't depend on Cr at all, which is incompatible with any DAB-type link.)

Some have objected to Invariance on the grounds that it contradicts certain heuristics for how to measure degrees of desire. [Stefánsson 2014], for example, points out that, by Averaging, the desirability of any proposition with probability 1 equals the desirability of the tautology. One might think that the tautology is neither desirable nor undesirable: it is neither "good news" nor "bad news", for it is no news at all. This suggests using the tautology to define a zero point for desirability:

$$V(\top) = 0. \qquad \text{(Normalization)}$$

Since $V^A(A) = V^A(\top)$, Invariance and Normalization (applied to V^A) together would imply that $V(A) = 0$, for all A. So if we accept Normalization, we must reject Invariance. To preserve Normalization after conditioning, we have to rescale the desirabilities, as suggested in [Bradley 1999]:

$$V^E(A) = V(A \wedge E) - V(E). \qquad \text{(Rescaling)}$$

A good discussion of the underlying problem can be found in [Jeffrey 1977]. Invariance is motivated by Stability. Stability says that conditionalizing on some information doesn't change an agent's basic desires. This is intuitively plausible. But so is Normalization, and we can't have both. Jeffrey argues that the choice is a philosophically insignificant matter of scaling. If we hold fixed the basic desire function $\mathrm{V_b}$, we measure the new desirabilities on the same scale as the old desirabilities, with the same zero. If instead we hold fixed the neutrality of the tautology, we measure the new desirabilities on a new scale in which the zero is moved to the previous desirability of the proposition that was learned. There is no deeper significance in this question about the choice of zero.

As one might expect, the choice makes no substantive difference to the refutation of DAB. Here is

a Lewis-style argument using Bradley's rescaled measure of V^A:

1.	$V(A) = \mathrm{Cr}(\mathring{A})$	[DAB]
2.	$V^A(A) = V(A) - V(A)$	[Rescaling]
3.	$V^A(A) = \mathrm{Cr}^A(\mathring{A})$	[DAB for Cr^A, V^A]
4.	$\mathrm{Cr}(\mathring{A}/A) = 0$	[from 2 and 3.]

The conclusion is no more acceptable than Independence.

There may, however, be a deeper reason to reject Invariance. Let w be a world in which rape and murder are (a) pervasive and (b) good. Is this a good world? Is it desirable? I'd say it is not. I'm confident in systems of norms that deem all rape-and-murder worlds bad, never mind whether these worlds regard themselves as good. For me, $V(w)$ and $\mathrm{Cr}(\mathring{w})$ are low. Conditional on w, however, w is good. This suggests that $V^w(w)$ is high. It would follow that $V^w(w) \neq V(w)$. We seem to have a counterexample to Invariance, and to Stability. More clearly, we seem to have a counterexample to Fetishism: since w is good according to itself, Fetishism would demand that $V(w) = 1$. In fact, I claim that $V(w)$ is low.

The triviality arguments from section 4 all assumed Fetishism. Lewis's argument assumes Invariance. Have we found a way out of the trouble?

Sadly, no. For one thing, the present suggestion doesn't dissolve the static counterexamples from the previous section. Besides, we still get triviality.

Let's first think about what to put in place of Fetishism. The above line of thought suggests that we may determine the desirability of a world not by looking at what the world itself says about its goodness, but by our actual, unconditional beliefs about what is good.

Let's assume, as before, that a complete hypotheses about goodness v assigns a goodness score $v(w)$ to each world. Instead of Fetishism, we might then have the following link between basic desire and beliefs about goodness:[16]

$$V_b(w) = \sum_v \mathrm{Cr}(v)v(w). \qquad \text{(Alignment)}$$

Informally, this says that the agent's basic desires are aligned with their normative beliefs. The degree to which they desire a given world depends on what they believe about its goodness. In section 9, I'll explain in more detail how this escapes the charge of fetishism. For now, I only want to show that it doesn't escape triviality.

Combining Alignment with Averaging, we get

$$V(A) = \sum_w Cr(w/A) \sum_v Cr(v)v(w).$$

By DAB, this yields

$$\mathrm{Cr}(\mathring{A}) = \sum_w \mathrm{Cr}(w/A) \sum_v \mathrm{Cr}(v)v(w). \qquad \text{(BACB*)}$$

Now $\mathrm{Cr}(\mathring{A})$ doesn't just depend on the distribution of credence within A. Outside A, however, $\mathrm{Cr}(\mathring{A})$

16 In Alignment, 'v' ranges over maximally specific normative propositions; I assume that each such proposition deter-

only depends on the credence distribution over normative matters. And that's not enough. For example, we still run into all the triviality problems for agents who are certain of a particular system of norms. Alternatively, we can still derive Independence as long as A is purely descriptive, and we can still falsify these instances of Independence by moving around credence.

7 Desire as Conditional Belief

[Price 1989] offers an attractive suggestion for how to avoid Lewis's refutation of DAB. Price points out that the calculation of V involves conditional credence, while DAB links V to unconditional judgements about goodness. The region outside A plays a role for the assessments of A's goodness, but not for $V(A)$. To fix this, Price suggests that we should equate $V(A)$ with the conditional credence that A is good, given that A is true:

$$V(A) = \mathrm{Cr}(\mathring{A}/A). \tag{DACB}$$

(The more plausible generalization, for more than two levels of goodness, would say that $V(A) = \sum_x \mathrm{Cr}(g(A) = x/A)$.)

Lewis discusses this proposal in [Lewis 1996]. In response, he gives a somewhat complicated proof by which, he says, DACB is "unmasked" as a thesis he calls "Desire by Necessity". He concludes that it is "the wrong kind of anti-Humeanism". The whole discussion is puzzling (as highlighted, for example, in [Hájek 2015]). Earlier in the paper, he describes Desire by Necessity as "comparatively simple and unproblematic". What's wrong with it?

Let's investigate. Desire by Necessity is the hypothesis that there are some propositions that everyone must desire. In its simplest form, it says that there is a fixed proposition G such that

$$V_b(w) = \begin{cases} 1 \text{ if } w \in G \\ 0 \text{ if } w \notin G \end{cases} \tag{DBN}$$

We've seen this before: Fetishism is an instance of DBN. In this case, G comprises all worlds that are good according to themselves. Lewis argued that anti-Humeans should accept Fetishism ([Lewis 1988: p.332]), and he pointed out that it is entailed by DAB ([Lewis 1996: sec.6]). So he could hardly have thought that it would be problematic if the anti-Humean had to accept DBN. His objection, I think, is that the anti-Humean should say *more* than DBN.

Early in [Lewis 1996], Lewis distinguishes two kinds of anti-Humeanism. One rejects Hume's *liberalism about desire*. According to Hume, there are no substantive rationality constraints on basic desires. ("It is not contrary to reason to prefer the destruction of the whole world" etc.) The other kind of anti-Humeanism rejects Hume's claim that *desires are independent of beliefs*: that there are no necessary connections between what an agent rationally believes and what they desire.

DBN falls in the first class. It rejects Humean liberalism. Lewis finds the proposed constraint on basic desire is implausible, at least if understood as an analytic truth. But this view is obviously compatible with decision theory. That's the sense in which DBN is "simple and unproblematic".

mines a goodness score for each world; I use '$v(w)$' for the score of world w under v.

But DBN has nothing to say about goodness, as a property of propositions. It doesn't link beliefs about the goodness of A to desires towards A. And that's what we wanted. We were looking for a kind of anti-Humeanism that appears to contradict Hume's second thesis, on the independence of belief and desire. If DACB reduces to DBN, this suggests that it is "the wrong kind of Humeanism". It is not what we were looking for.

I find this argument (if it was Lewis's) a little elusive. In Jeffrey's decision theory, Averaging ensures that V is determined by V_b and Cr. DBN fixes V_b. But V_b and Cr fix V. Any further principle about how V relates to Cr is either redundant or inconsistent with Averaging. That is, if the anti-Humean wants to remain consistent with Averaging, one might worry that their postulated belief-desire link *can't* go beyond DBN.

Still, I think Lewis was onto something. Remember that Fetishism/DBN and Averaging got us $V(A) = \mathrm{Cr}(G/A)$. DACB says that $V(A) = \mathrm{Cr}(\mathring{A}/A)$. An easy way to uphold DACB is therefore to identify $\mathring{A}$ with G. The halo would be a constant function, mapping every proposition A to the same proposition G. But if that's the proposal, $\mathring{A}$ could hardly be understood as the hypothesis that A is good. The hypothesis that donating to charity is good is not identical to the hypothesis that torturing kittens is good!

The anti-Humean idea we are trying to formalize assumes that some propositions (acts, events, etc.) have a special property of goodness that makes them desirable: a belief that donating to charity is good would be connected to a desire to donate to charity, a belief that torturing kittens is good would be connected to a desire to torture kittens, etc. DBN doesn't give us any such link. It links desires to beliefs, but the relevant beliefs aren't about what is good. DACB *appears* to give us such a link. But if the halo is a constant function, it does not, for then $\mathring{A}$ isn't a hypothesis about A at all. We don't get a link between the desirability of A and beliefs about the goodness of A. We really don't get what we were looking for.

Now, DACB doesn't require identifying $\mathring{A}$ and G. Given Fetishism, however, it requires identifying $\mathring{A}$ and G *within* A: $\mathring{A} \wedge A$ must equal $G \wedge A$. If it doesn't, we could falsify DACB by moving around credence. Lewis shows that this is true even without the assumption of Fetishism: if DACB holds unrestrictedly, he shows, there must be a fixed proposition G such that for all A, $\mathring{A} \wedge A = G \wedge A$.[17] And this is bad enough. Remember that the only part of Cr that matters for $V(A)$ in Jeffrey's framework is the part of Cr that lies within A. So DACB implies that while there might be a difference between believing that donating to charity is good and believing that torturing kittens is good, the difference is irrelevant to the extent to which one might desire these events.

All this is somewhat obscured by a quirk in Lewis's setup. Lewis introduces his anti-Humean target as an existential hypothesis. The anti-Humean is supposed to hold that *there is a halo function* for which a DAB-like connection holds, as a matter of analytic necessity. If there is such a function, Lewis suggests, it would be natural to interpret $\mathring{A}$ as the hypothesis that A is (objectively) good. In effect, Lewis assumes that the connection to desire defines the concept of goodness: goodness is whatever property – if any – is such that merely believing that something has it makes that thing desirable.[18] But that's not true. We know more about goodness (and other normative properties). We

17 G is the proposition $\mathring{\top}$ that the tautology is good. By Lewis's "Downward Lemma", DACB entails that $V(w) = 1$ if $w \in \mathring{\top}$ else $V(w) = 0$. The "Upward Lemma" then shows that $\mathring{A} \wedge A = \mathring{\top} \wedge A$.

18 In fact, Lewis only suggests one direction of the analysis: *if* there is a halo function that satisfies DAB *then* $\mathring{A}$ can be

know, for example, that the hypothesis that donating to charity is good is different from the hypothesis that torturing kittens is good. We know that some things are better than others. If the halo means goodness, it is not a constant function.

The counterexamples to DAB-type principles from section 5 touch on this point. Remember the second greenhouse gas example. Like DAB, DACB entails that $\mathrm{Cr}(A \wedge B / \neg \mathring{A} \wedge \mathring{B}) = 0$.[19] But in the greenhouse gas case, this is intuitively false. We know that A is bad and B is good, even though there's a small chance that both are true. We know more about goodness than what's entailed by DAB or DACB, just as we know more about indicative conditionals than what's entailed by Stalnaker's Thesis.

Remember also the rape-and-murder world w that regards itself as good. I argued that this world is not good, and that it isn't desirable. This directly contradicts DACB, for w is certainly good conditional on itself: $\mathrm{Cr}(\mathring{w}/w) = 1$. Here, too, we have a judgement about goodness that contradicts DACB.

We have yet to find a way out.

8 Objective value, subjunctive desire

Perhaps we should drop Averaging. Recall how I motivated this assumption. If a proposition divides into more and less desirable parts, how strongly should it be desired? Averaging says that we should weight the parts by their probability, conditional on the proposition. This is a sensible way to determine desirability. But it is not the only one. Intuitively, it measures how strongly we desire that the proposition *is* true, as a hypothesis about the actual world. A different measure looks at what *would* be the case if the proposition *were* true.

An example. Werner Heisenberg led the Nazi nuclear weapon program, which failed. If Heisenberg knew how to build nuclear weapons, he must have kept it secret, subtly sabotaging the program. Personally, I would like to learn that this is true, as I'd like to think of Heisenberg as a good person. So I desire that Heisenberg knew how to build nuclear weapons.

People who know more about this issue than me generally agree that Heisenberg did *not* know how to build nuclear weapons, and that he would not have kept such knowledge secret. In light of this, it would have been terrible if Heisenberg had known how to build nuclear weapons. I'm glad that he didn't. In one sense, I desire that Heisenberg knew how to build nuclear weapons; in another, I don't.

Formally, the difference between the two kinds of desire lies in how they weigh the subregions of the relevant proposition. The first, *indicative* kind of desire uses Averaging. The second, *subjunctive* kind of desire, uses a subjunctive form of averaging: if A divides (exhaustively and exclusively) into A_1 and A_2, then

$$U(A) = U(A_1)\,\mathrm{Cr}(A_1 // A) + U(A_2)\,\mathrm{Cr}(A_2 // A). \qquad \text{(Subjunctive Averaging)}$$

understood as the proposition that A is objectively good. As [Weintraub 2007] points out, the real force of Lewis's argument turns on the converse direction: if there is no such halo, it seems that we can't make sense of objective goodness.

19 Let Cr', V' be Cr, V after conditioning on $A \wedge B$. By Averaging, $V'(A) = V'(B)$. So by DABC, $\mathrm{Cr}'(\mathring{A}) = \mathrm{Cr}'(\mathring{A}/A) = V'(A) = V'(B) = \mathrm{Cr}'(\mathring{B}/B) = \mathrm{Cr}'(\mathring{B})$. If there was any region in $A \wedge B$ where $\mathring{A}$ was false and $\mathring{B}$ true, we could falsify this equality by concentrating Cr' on that region. So $\mathring{A}$ and $\mathring{B}$ coincide within $A \wedge B$ and $\neg \mathring{A} \wedge \mathring{B}$ lies entirely outside $A \wedge B$.

Here, U stands for subjunctive desire, and $\mathrm{Cr}(A_1//A)$ is the credence that A_1 would be the case if A were the case. The double slash is defined by (generalised) *imaging*: $\mathrm{Cr}(A//B)$ is the probability that results when moving the probability mass of any non-B world to the "closest" B-worlds, relative to a certain measure of closeness. (For more details, see e.g. [Joyce 1999: ch.6].)

Both U and V arise from the same basic value function V_b:

$$U(A) = \sum_w V_b(w)Cr(w//A).$$

Now, we've tried unsuccessfully to link beliefs about goodness to indicative desire. What if we switch to subjunctive desire?

Instead of DAB, we might try:

$$U(A) = \mathrm{Cr}(\mathring{A}). \qquad \text{(Subjunctive DAB)}$$

Or, for more than two levels of goodness:

$$U(A) = \sum_x x\,\mathrm{Cr}(g(A) = x). \qquad \text{(Subjunctive DAE)}$$

Assuming Fetishism, Subjunctive DAB yields a subjunctive version of BACB:

$$\mathrm{Cr}(\mathring{A}) = \mathrm{Cr}(G//A). \qquad \text{(Subjunctive BACB)}$$

Does this share the fate of BACB?

It depends. Suppose the imaging operation is *sharp*, so that it never moves a single world to multiple A-worlds. As [Lewis 1976] showed, one can then define a "Stalnaker conditional" $A > C$ for which

$$Cr(A > C) = Cr(C//A).$$

$A > C$ is defined to be true at w iff the "closest" A-world from w is a C-world. If we now read $\mathring{A}$ as $A > G$, we get Subjunctive DAB.[20]

Why doesn't Subjunctive BACB run into the problem for BACB? Because it allows $\mathring{A}$ to be information-independent, if we have sharp imaging. $A > G$ makes a cut through the space of worlds. For each world, it is either true or false that the closest A-world is a G-world, irrespective of any information parameter.

So here, finally, we seem to have a way out. We can link beliefs about the goodness of propositions to the agent's subjunctive desire towards that proposition. And maybe we should have done this all along, given that subjunctive desire is – arguably – more directly related to motivation than indicative desire. (See, e.g., [Lewis 1981], [Skyrms 1982], [Joyce 1999].) Problem solved?

Not quite. I'll mention three worries.

First. It is doubtful that there is always a unique "closest" world at which a proposition is true. The standard counterexample (endorsed by Lewis) involves an indeterministic coin flip. In an indeterministic world, the laws of nature and the total past may not determine how a coin will land. Suppose w is such a world, and some indeterministic coin in w isn't flipped. It seems that nothing in w settles

20 This appears to have been first pointed out in Jessica Collins's PhD thesis from 1991.

how the coin *would* have landed if it *had* been flipped: not the past and the laws, evidently, and the present or future of w don't seem to provide the missing information.

If there are ties in closeness, we're back in trouble. For a simple example, suppose Cr is concentrated on a single world w for which there are two closest A-worlds, one in G, the other outside G. (Perhaps it would be good if a certain coin that's never flipped had landed heads, and bad if it had landed tails.) $\mathrm{Cr}(\mathring{A})$ must be 0 or 1, because Cr is concentrated on a single world; but $\mathrm{Cr}(G//A)$ will be strictly between 0 and 1. So Subjunctive BACB fails.[21]

Second. The above construction assumes that we can simply define $\mathring{A}$ as $A > G$. But remember that our goal was to link *beliefs about goodness* to desire, so the halo had better respect what we already know about goodness. For example, if we wanted to link an information-sensitive concept of goodness to desire – a sense in which, for example, it is best to block neither shaft in the miners puzzle, – then Subjunctive DAB won't give us what we want. Or suppose we thought that goodness was primitive, not definable in terms of anything else. Then, again, the present approach doesn't deliver, as it assumes that 'A is good' can be defined as $A > G$.

Perhaps these problems are not too serious. Arguably, the lesson of the triviality results is that apparent beliefs about information-sensitive goodness can't be adequately modelled in a probabilistic framework. And perhaps it is a conceptual prejudice to think that goodness is undefinable. But there's a more serious version of the problem.

Our derivation of Subjunctive DAB assumed Fetishism. So it faces the rape-and-murder objection from section 6. In fact, that objection is even stronger here, for we can now assume that we give credence zero to the rape-and-murder world w that regards itself as good. Subjunctive conditional probability is defined even for propositions with probability 0. Specifically, $\mathrm{Cr}(\mathring{w}//w) = 1$. By Subjunctive BACB, it follows that $\mathrm{Cr}(\mathring{w}) = 1$. But w is not good!

Unfortunately, we can't easily drop Fetishism. Suppose we trade it for Alignment. Combining Subjunctive DAB with Alignment, we get

$$\mathrm{Cr}(\mathring{A}) = \sum_w \mathrm{Cr}(w//A) \sum_v \mathrm{Cr}(v)v(w). \qquad \text{(Subjunctive BACB*)}$$

This makes $\mathring{A}$ information-dependent, even if imaging is sharp.

The trouble is a little hard to visualize, but here is one way to see it. To begin, if $\mathring{A}$ is a region in logical space, we must have $w \in \mathring{A}$ iff $\mathrm{Cr}(\mathring{A}/w) = 1$, provided $\mathrm{Cr}(w) > 0$. By Subjunctive BACB*, $\mathrm{Cr}(\mathring{A}/w) = 1$ for precisely those worlds w that assess their closest A-world as good. Let S be the set of these worlds. Among worlds with positive probability, we must have $\mathring{A} = S$. Now there should be cases where S contains some A-world. (For A and $\mathring{A}$ shouldn't always be inconsistent with one another.) If so, let w_1 and w_2 be two worlds outside S that disagree about the goodness of some A-world in S. By Subjunctive BACB*, $\mathrm{Cr}(\mathring{A})$ depends on the relative distribution of credence over w_1 and w_2: we can change $\mathrm{Cr}(\mathring{A})$ by shifting credence between w_1 and w_2, assuming both have positive credence. But $\mathrm{Cr}(S)$ is unaffected by such shifts, as w_1 and w_2 are outside S. Contradiction.

All this was part of my second complaint: we can't combine Subjunctive DAB with independently

21 If you object to a credence function that is concentrated on a single world, let A be the hypothesis that the indeterministic coin is flipped; assume you're confident that it would be good if it landed heads and bad if it landed heads. Then let Cr be your credence conditional on $\mathring{A}$. We have $\mathrm{Cr}(\mathring{A}) = 1$, but $\mathrm{Cr}(G//A) = 0.5$.

plausible constraints on the meaning of the halo. My third complaint is simpler. It relates to the question whether what we'd get is "the right kind of anti-Humeanism".

No sensible Humean ever denied that beliefs can motivate, if they are accompanied by suitable desires. If you have an intrinsic desire to be healthy, and you believe that you will be healthy only if you eat carrots, you will be motivated to eat carrots. Holding fixed your intrinsic desires, we'll obviously find a link between your beliefs and your (non-intrinsic) desires. This is not a serious violation of Humean independence. Your belief about carrots is motivating, alright, but its motivational force comes from your desire for health.[22] Isn't the connection between normative beliefs and motivation supposed to be different? Normative beliefs are supposed to be *intrinsically motivating*, as the saying goes. Subjunctive DAB doesn't deliver this. On the model we've developed, all agents have a basic desire for G. A belief that $A > G$ is therefore motivating, alright, but its motivational force appears to be extrinsic, derived from the desire for G.

This worry will become a little clearer after we've clarified what it means for something to be an intrinsic desire.

9 Intrinsic value, intrinsic desire

According to classical utilitarianism (as well as many of its rivals), happiness is an *intrinsic* or *final good*. What does this mean?[23] The proposition that, say, you are happy isn't uniformly good. Worlds where you are happy and everyone else is miserable are worse than worlds where you are happy and so is everyone else. Suppose your happiness is, for contingent reasons, negatively correlated with that of other people, so that it's likely that others are miserable if you are happy. It wouldn't follow that your happiness isn't intrinsically good. Intrinsic goodness doesn't involve averaging over subregions. Informally, it's a matter of what a proposition contributes to the total goodness of a world. To say that your happiness is an intrinsic good is to say that it makes a positive contribution: all else equal, worlds in which you are happy are better than worlds in which you are unhappy.

An analogous concept of desire is studied in utility theory. So far, we've assumed that an agent's basic desires are represented by the assignment V_b that assigns a desirability score to individual worlds. Realistically, however, the desirability of a world is determined by what happens at that world, and most of the things that happen are irrelevant. The truly basic desires pertain to *aspects* of worlds. I call these *intrinsic desires*.

Formally, we can represent an agent's intrinsic desires by a new function V_i that assigns a value to certain aspects of the world – that is, to certain propositions. Intuitively, these are the aspects the

22 This might be understood literally, as a causal claim. A corrsponding construal of the Humean theory posits beliefs and desires as separate causal nodes in the genesis of rational action: the agent would be assumed to have a "belief module" and a "desire module" that act together to cause behaviour. But there is no good reason to equip agents with a special desire module if their basic desires never change. One could directly build the agent's goals into the decision algorithm. (Compare [Pettit and Price 1989].) I therefore prefer to think of the Humean theory not as the speculative postulation of two psychological modules, but as the more abstract hypothesis that a full psychological explanation of agency must cite both beliefs and desires, whether or not these are implemented in separate modules.

23 The following explication is closer to what [Korsgaard 1983] calls 'final goodness' than what she calls 'intrinsic goodness'. But 'final' contrasts with 'instrumental', and I find it odd to say that if your happiness is a final good and the outcome of a die toss is irrelevant then *you are happy and the die lands six* is instrumentally good.

agent intrinsically cares about. They often come in families. If you have an intrinsic desire to be happy, and happiness comes in degrees, then your V_i function assigns a score to each proposition in the family $\{H = x\}_x$ that specifies your degree of happiness. If you also have an intrinsic desire that your pet rabbit is happy, your V_i function assigns another score to each proposition in the family $\{H_r = x\}_x$ that specifies your rabbit's degree of happiness. Your "basic desire" for a world $V_b(w)$ is determined by adding up the scores of its aspects:

$$V_b(w) = \sum_{A:w\in A} V_i(A). \qquad \text{(Additivity)}$$

Here, A ranges over the propositions in the domain of V_i.[24] If all you care about is the happiness of you and your rabbit, $V_b(w)$ is the sum of the score you assign to your happiness in w and the score you assign to your rabbit's happiness in w.

Additivity assumes that the agent's preferences over the different aspect families are independent. For example, if you prefer to be happy if your rabbit is happy, but not if your rabbit is sad, we couldn't determine your overall desire towards a world by adding up separate scores for your happiness and your rabbit's happiness. In that case, however, you wouldn't have an *intrinsic* desire for the two aspects. You might have an intrinsic desire for *you and your rabbit to be happy together*, but you don't have an intrinsic desire for you to be happy. Intrinsic desires are, by definition, always independent.[25] (We can allow for agents with extreme bouletic holism who intrinsically care only about a single, very fine-grained family of propositions.)

Note that $V_i(A)$ and $V(A)$ generally come apart, even if they are both defined. $V(A)$ is the "overall" desirability of A, taking into account what else is likely to be the case if A is true. $V_i(A)$ is the intrinsic desirability of A, it's the contribution made by A to the overall desirability of a world.

Intrinsic desire can't be directly linked to beliefs about "overall goodness" – the information-sensitive kind of goodness that takes into account the probability of subregions. But we could try to link intrinsic desire to beliefs about *intrinsic* goodness.

What might this link look like? There's a "de dicto" option and a "de re" option. The de dicto option posits a single intrinsic desire: that whatever is intrinsically good be the case. The de re option posits multiple intrinsic desires towards all the things that are believed to be intrinsically good.

The de dicto route leads to Fetishism. If a world w says that such-and-such things are intrinsically good, and these things are the case at w, an agent with a de dicto desire for goodness will desire w. We don't really get a *connection* between intrinsic desire and beliefs about intrinsic goodness. We get "the wrong kind of anti-Humeanism". Let's see if we can do better with the de re option.

24 My simple formulation of Additivity assumes that no proposition occurs in more than one of the families of propositions that the agent intrinsically cares about. If this condition isn't satisfied, we would have explicitly sum over each family.

25 What I'm describing here is a version of *multi-attribute utility theory*, ported to Jeffrey's framework, where the objects of desire are propositions rather than "outcomes". My "families of aspects" correspond to the "attributes" of outcomes. In utility theory, the numerical representation of (basic) desire is assumed to be derived from a preference order over outcomes. [Debreu 1960] proved that if this order is complete, transitive, continuous, and strongly separable then it can be represented by an additive utility function over the outcomes and its attributes. Strong separability is the independence requirement. It means that if two outcomes differ in some of the attributes and not in others then their ranking depends only on the attributes in which they differ. Continuity requires that the different attributes are commensurable. See [Krantz et al. 1971] for the standard textbook on multi-attribute utility theory and its mathematics.

We have to assume that normative beliefs pertain to hypotheses about intrinsic goodness. Let's assume, then, that a complete hypothesis about goodness v determines an assignment of an intrinsic goodness score v_i to certain aspects of worlds, so that a world's total score $v(w)$ is given by the sum of the scores of its aspects:[26]

$$v(w) = \sum_{A:w\in A} v_i(A). \qquad \text{(Additivity of Goodness)}$$

(As in section 6, I assume that v_i ranges over the propositions in certain families, each of which represents a (separable) dimension of value. As above, we can allow for extreme value holism, where there is only a single, fine-grained family.)

We might now suggest that the intrinsic desirability of a proposition equals its expected intrinsic goodness:

$$V_i(A) = \sum_v \mathrm{Cr}(v)v_i(A). \qquad \text{(Intrinsic DAE)}$$

That is, if you're sure that A is intrinsically good to degree x, you will have an intrinsic desire for A of strength x; if you're undecided whether the intrinsic goodness of A is x_1 or x_2, your intrinsic desire for A is the average of x_1 and x_2; and so on.

In the unrealistic case where all goodness scores are 0 or 1, Intrinsic DAE reduces to

$$V_i(A) = \mathrm{Cr}(\dot{A}), \qquad \text{(Intrinsic DAB)}$$

where the dot expresses intrinsic goodness. We have a form of DAB, but for intrinsic desire and beliefs about intrinsic goodness.

As stated, intrinsic DAE assumes that all candidate value functions v assign an intrinsic goodness score to A. To fix this, we could stipulate that the sum only ranges over those v that do assign an intrinsic score to A. Equivalently, we could stipulate that if $v(A)$ is undefined then it is treated as 0. I'll adopt this second convention. Thus $V_i(A)$ is 0 if no candidate value function assigns an intrinsic goodness score to A, meaning that A makes not contribution to the desirability of worlds.

What does Intrinsic DAE imply about the basic desirability function V_b? By Additivity,

$$V_b(w) = \sum_{A:w\in A} V_i(A).$$

Combining this with Intrinsic DAE, we get

$$V_b(w) = \sum_{A:w\in A} \sum_v \mathrm{Cr}(v)v(A) = \sum_v \mathrm{Cr}(v) \sum_{A:w\in A} v(A).$$

By Additivity of Goodness,

$$v(w) = \sum_{A:w\in A} v(A).$$

26 I use the same overloading of the variable 'v' as in section 6: 'v' ranges over maximally specific normative propositions, but also stands for the value function determined by any such proposition, so that '$v(w)$' denotes the score of world w under v.

So

$$V_b(w) = \sum_v \mathrm{Cr}(v)v(w).$$

We get Alignment.

By inverting this line of thought, we can almost derive Intrinsic DAE from Alignment and the Additivity principles: Alignment and Additivity of Goodness together imply that

$$V_b(w) = \sum_v \mathrm{Cr}(v) \sum_{A:w\in A} v(A) = \sum_{A:w\in A} \sum_v \mathrm{Cr}(v)v(A)$$

This has the form of Additivity. It represents the agent's basic desires towards a world as arising from intrinsic desires towards all the aspects of the world which they think might have an intrinsic goodness value. The intrinsic desire function V_i implicit in this representation is given by Intrinsic DAE:

$$V_i(A) = \sum_v \mathrm{Cr}(v)v(A).$$

(I say that this is *almost* a derivation of DAE because one might adopt a "realist" understanding of intrinsic desire, on which showing that an agent's basic desires can be represented as arising from certain intrinsic desires is not enough to show that the agent genuinely has those intrinsic desires.)

To get a better sense of what Intrinsic DAE and DAB mean, return once more to the rape-and-murder world w according to which rape and murder are good.

Suppose – reasonably – that most of our credence goes to normative theories according to which rape and murder are intrinsically bad. Since $V_b(w) = \sum_{A:w\in A} \sum_v \mathrm{Cr}(v)v(A)$, it follows that $V_b(w)$ is low. We don't desire that w is actual. But if we were to learn that w is actual, we would learn that rape and murder are good. $V_b(w)$ would become high.

We have to be careful when we think of V as a measure of "news value". We must distinguish our actual desire towards a proposition from what we would come to desire if we were to learn that proposition. On the present model, learning w would change our desire towards w. As a mere possibility, w is "bad news". But if we were to receive w as actual news, we would no longer see it as bad news. The news that w is actual would *make* it good news.

Earlier, I sometimes considered a pair of updated attitudes Cr^E, V^E that are supposed to arise from Cr, V by conditionalizing on E. It's clear what this means for Cr^E, but what does it mean for V^E? It depends on how we read V^E.

On one reading, $V^E(A)$ is the *revised* desirability of A, after having incorporated the information E. On this reading, Invariance and Stability fail. On another reading, $V^E(A)$ is the desirability of A *conditional* on E, defined by

$$V^E(A) = \sum_w V_b(w)Cr^E(w/A).$$

Here we hold fixed the basic value function V_b and merely conditionalize Cr to Cr^E.[27] On this reading, Invariance and Stability hold, but we can't assume that V^E, Cr^E satisfy Intrinsic DAE, even if V, Cr

27 See [Bradley 1999], [Bradley 2017: ch.6], [Joyce 1999: ch.6], and [Joyce 2020] for more on conditional desirability and its theoretical role. My proposed definition follows Joyce; Bradley prefers a "re-scaled" version of the definition that makes $V^E(E)$ equal zero.

is a rational pair of attitudes and it is rational to conditionalize on E. For V^E doesn't process the information in E to revise the intrinsic desirability function, as would be required by DAE if the agent actually learned that E is true.

What does the present model say about the overall (non-intrinsic) desirability of regions in logical space? We already know from sections 6 what we get if we combine Alignment with Averaging:

$$V(A) = \sum_w \mathrm{Cr}(w/A) \sum_v \mathrm{Cr}(v)v(w).$$

In section 6, we explored setting $\mathrm{Cr}(\mathring{A})$ equal to $V(A)$, on this construal of $V(A)$. The resulting equality, BACB*, ran into triviality. It would make $\mathring{A}$ information-sensitive.

Now, we *do* have an information-sensitive concept of goodness. We can check how it might work.

Suppose we want to extend an objective value function v so that it gives a verdict about propositions that aren't intrinsically good or bad. Such a proposition may have better and worse parts. To give a sensible verdict, we must supply v with a probability measure over these parts. Let v^{Cr} be the value function induced by v and Cr so that

$$v^{Cr}(A) = \sum_w Cr(w/A)v(w).$$

We can now define a simple information-sensitive concept of goodness, g^{Cr}, by stipulating that $g^{\mathrm{Cr}}(A) = x$ is true at w iff $v_w^{\mathrm{Cr}}(A) = x$, where v_w is the objective value function according to w. The proposition expressed by $g^{\mathrm{Cr}}(A) = x$ depends on the information parameter Cr.

Now remember that when we assess the probability of a simple information-sensitive statement A, we seem to pass the very probability measure that is used for the assessment into A's information slot. We can therefore model our probability judgements about whether a proposition A is good to degree x, on an information-sensitive construal of goodness, by assuming that they evaluate $\mathrm{Cr}(g^{\mathrm{Cr}}(A) = x)$.

To illustrate, consider the miners puzzle. Here we have no significant normative uncertainty. All normative hypotheses v with positive credence, we may assume, assign intrinsic value to the life of each miner. For concreteness, let's say that $v(w)$ equals the number of surviving miners in w, for all live hypotheses v. If blocking neither shaft is sure to lead to the death of one miner, $g^{\mathrm{Cr}}(\textit{Block Neither})$ is known to equal 9. More interestingly, $g^{\mathrm{Cr}}(\textit{Block A})$ is known to equal 5, because g^{Cr} evaluates *Block A* by computing the Cr-weighted average of $v(w)$ for each world w inside *Block A*. $g^{\mathrm{Cr}}(\textit{Block B})$ is also known to equal 5. So we know that *Block Neither* is best. On the supposition that the miners are in shaft A, however, we evaluate the goodness of *Block A* and *Block Neither* relative to the updated credence Cr' that results from Cr by conditionalizing on the supposition that the miners are in shaft A. $g^{\mathrm{Cr}'}(\textit{Block Neither})$ is still 9, but $g^{\mathrm{Cr}'}(\textit{Block A})$ now becomes 10. We are sure that *Block Neither* is worse than *Block A*. By the same reasoning, we are sure that *Block Neither* is worse than *Block B*, on the supposition that the miners are in shaft B. Thus we arrive at the probabilistically incoherent judgement that *Block Neither* is best unconditionally, but not conditional on either of the two possibilities about the miners' location.

For a different type of illustration, consider a case of pure normative uncertainty. We're in a trolley problem, and we're undecided between two moral theories, v_1 and v_2. We can determine what the

two theories say about the option of flipping the switch, in light of our information Cr about the scenario, by computing $v_1^{\mathrm{Cr}}(\mathit{Flip})$ and $v_2^{\mathrm{Cr}}(\mathit{Flip})$. Let's say that $v_1^{\mathrm{Cr}}(\mathit{Flip}) = x_1$ and $v_2^{\mathrm{Cr}}(\mathit{Flip}) = x_2$.[28] Intuitively, x_1 is the goodness of *Flip* according to v_1. Indeed, $\mathrm{Cr}(g^{\mathrm{Cr}}(\mathit{Flip}) = x_1/v_1) = 1$. Likewise, $\mathrm{Cr}(g^{\mathrm{Cr}}(\mathit{Flip}) = x_2/v_2) = 1$. The *expected* goodness of *Flip*, in our state of indecision between v_1 and v_2, is

$$E[g^{Cr}(\mathit{Flip})] = x_1 Cr(v_1) + x_2 Cr(v_2).$$

We might think that this should match $V(\mathit{Flip})$. And it does! From Alignment and Averaging, we have

$$V(A) = \sum_w \sum_v \mathrm{Cr}(w/A)\,\mathrm{Cr}(v) v(w).$$

By definition of v^{Cr}, this yields

$$V(A) = \sum_v Cr(v) v^{Cr}(A).$$

In the example, this means that

$$V(\mathit{Flip}) = x_1 Cr(v_1) + x_2 Cr(v_2).$$

So $V(\mathit{Flip}) = E[g^{\mathrm{Cr}}(\mathit{Flip})]$. In general, $V(A) = E[g^{\mathrm{Cr}}(A)]$. We only run into trouble if we think that the "expected goodness of A" is the expectation of a genuine quantity, "the goodness of A". But '$g(A) = x$' is information-sensitive: there is no such thing as $g(A)$; there is only $g^{\mathrm{Cr}}(A)$.

We've just derived "moral hedging". To be clear, my aim is not to defend moral hedging. I only want to explain how one can link beliefs about goodness to desire, without running into trouble. If there is such a link, it's natural to expect some form of moral hedging: if your confidence that something is good affects your degree of desire, one would expect that being undecided whether something is good shows up in a middling degree of desire. In the model I have presented, Intrinsic DAE builds in a simple form of hedging (assuming that it is applied to a moral reading of 'v'). One could easily tweak Intrinsic DAE to get, for example, more risk-averse forms of hedging.

An important objection to moral hedging (raised, for example, in [Weatherson 2019]) is that it requires moral fetishism: an agent whose desires are determined by their beliefs about goodness must, it seems, only have one intrinsic desire: that the world be good. But good people aren't only, or even primarily, motivated by abstract considerations of goodness. "Good people care non-derivatively about honesty, the weal and woe of their children and friends, the well-being of their fellows, people getting what they deserve, justice, equality, and the like", as [Smith 1994: p.75] memorably put it.

If basic (and intrinsic) desires are insensitive to belief, as Jeffrey assumed, the objection is correct: your desires can then sway with your beliefs about goodness only if they aren't pinned down by

28 Where do these numbers come from, and why are they comparable? There are different ways to think about this. Here is one. First, we assume that each moral theory specifies a "betterness" order over combinations of intrinsically valuable aspects. Under somewhat idealized conditions, it follows from Debreu's theorem that any such order has an additive numerical representation that is unique up to positive affine transformations. To make the numbers comparable, we could now, for example, fix a common range: stipulate that the worst combination must be assigned 0, the best 1. Alternatively, we use a variant of *Normality* (from section 6) to stipulate that the mean of each measure must be zero and the variance one, relative to some fixed probability measure. See, for example, [MacAskill et al. 2020b] for more on this line of thought.

intrinsic desires towards honesty and justice. In the model I have proposed, however, the objection fails. The very heart of my proposal is that people have a range of intrinsic desires that sway with their beliefs about goodness.

10 Conclusion

We've studied how normative beliefs might be connected to desire and motivation, in a decision-theoretic framework. We've looked at two types of agent, roughly corresponding to Smith's [1994] distinction between agents who care about morality "de dicto" and agents who care about it "de re".

The de dicto type can be easily modelled in Jeffrey's decision theory, by adding the assumption of Fetishism (a version of what Lewis calls "Desire by Necessity"). But the resulting model has serious problems. It is not only vulnerable to Smith's charge of fetishism. It also fails to establish a link between beliefs about the normative status of specific propositions (say, donating to charity) and a desire towards these propositions.

A more adequate model requires an adjustment to Jeffrey's theory, allowing basic desires to vary with the agent's normative beliefs. We still have to tread carefully to escape the triviality results, which show that there is no simple connection between beliefs about overall goodness and overall desire. The heart of these results is that the overall goodness of a proposition depends on the probability of its parts: "overall goodness" is an information-sensitive concept, and there is a general problem – a real problem – of fitting such concepts into a Bayesian framework. We could find the needed connection between belief and desire by focusing on beliefs about *intrinsic* goodness, which can be linked to intrinsic desires.

On reflection, it should have been clear from the outset that our anti-Humean endeavour requires departing from Jeffrey's assumption that the basic desire function V_b is insensitive to belief. Remember that V and U are determined by V_b and Cr. If we hold fixed V_b, there is no degree of freedom left. Any connection between Cr and V or U is either entailed by the Averaging rules or incompatible with these rules. DACB turned out to be redundant; DAB is incompatible; Subjunctive DAB is redundant with sharp imaging, incompatible otherwise. But we weren't looking for a *redundant* connection between belief and desire! We wanted to find a psychological model in which beliefs about goodness are non-trivially linked to desire. Our anti-Humean thesis should not be entailed by standard decision theory. The "right kind of anti-Humeanism" should make a substantive further claim – a claim like Intrinsic DAE, about which Humeans and anti-Humeans might now have a debate.

Fortunately, the departure from Jeffrey's theory is minimal. It is no part of folk psychology that basic desires must be insensitive to belief. Decision theory need not be Humean.

A formal model of how normative beliefs can motivate does not, of course, answer the motivational challenge to moral realism. It merely gives a precise shape to the challenge: if v is a hypothesis about a mysterious extra dimension of reality, why should our intrinsic desires be aligned with our credence in v, as expressed by Intrinsic DAE?

Antirealists are not entirely safe from the challenge. It is a Moorean fact that we have normative beliefs, and that they are linked to desire and motivation. The antirealist needs to explain how $\mathrm{Cr}(v)$ should be understood, and why it would be aligned with intrinsic desire.

References

Ernest W. Adams [1975]: *The Logic of Conditionals*. Dordrecht: D. Reidel

Andrew Bacon [2015]: "Stalnaker's Thesis in Context". *The Review of Symbolic Logic*, 8(1): 131–163

Jonathan Bennett [2003]: *A Philosophical Guide to Conditionals*. New York: Oxford University Press

Richard Bradley [1999]: "Conditional Desirability". *Theory and Decision. An International Journal for Multidisciplinary Advances in Decision Science*, 47(1): 23–55

— [2012]: "Multi-Dimensional Possible-World Semantics for Conditionals". *The Philosophical Review*, 121: 539–571

— [2017]: *Decision Theory with a Human Face*. 2017

Richard Bradley and Christian List [2009]: "Desire-as-Belief Revisited". *Analysis*, 69(1): 31–37

Richard Bradley and H. Orri Stefánsson [2016]: "Desire, Expectation, and Invariance". *Mind*, 125(499): 691–725

John Broome [1991]: "Desire, Belief and Expectation". *Mind*, 100(2): 265–267

Alex Byrne and Alan Hájek [1997]: "David Hume, David Lewis, and Decision Theory". *Mind*, 106: 411–728

Ivano Ciardelli and Adrian Ommundsen [2024]: "Probabilities of Conditionals: Updating Adams". *Noûs*, 58(1): 26–53

J. Collins [1988]: "Belief, Desire, and Revision". *Mind*, 97(387): 333–342

— [2015]: "Decision Theory after Lewis". In *A Companion to David Lewis,* chapter 28. John Wiley & Sons, Ltd, 446–458

Steven Daskal [2010]: "Absolute Value as Belief". *Philosophical Studies*, 148(2): 221–229

Gerard Debreu [1960]: "Topological Methods in Cardinal Utility Theory". In K Arrow, S Karlin and P Suppes (Eds.) *Mathematical Methods in Social Sciences,* Stanford University Press, 16–26

Dorothy Edgington [1995]: "On Conditionals". *Mind*, 104(414): 235–329

Branden Fitelson [2015]: "The Strongest Possible Lewisian Triviality Result". *Thought: A Journal of Philosophy*, 4(2): 69–74

Allan Gibbard [1981]: "Two Recent Theories of Conditionals". In William Harper, Robert C. Stalnaker and Glenn Pearce (Eds.) *Ifs,* Reidel, 211–247

Simon Goldstein and Paolo Santorio [2021]: "Probability for Epistemic Modalities". *Philosophers' Imprint*, 21(33)

Alex Gregory [2021]: *Desire as Belief: A Study of Desire, Motivation, and Rationality*. Oxford University Press

Alan Hájek [2015]: "On the Plurality of Lewis's Triviality Results". *A companion to David Lewis*: 425–445

Alan Hájek and Philip Pettit [2004]: "Desire Beyond Belief". *Australasian Journal of Philosophy*, 82(1): 77–92

Peter Hawke and Shane Steinert-Threlkeld [2021]: "Semantic Expressivism for Epistemic Modals". *Linguistics and Philosophy*, 44(2): 475–511

Richard Jeffrey [1965]: *The Logic of Decision*. New York: McGraw-Hill

— [1977]: "A Note on the Kinematics of Preference". *Erkenntnis*, 11(1): 135–141

James Joyce [1999]: *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press

— [2020]: "Conditional Desirability: Comments on Richard Bradley's Decision Theory with a Human Face". *Synthese*

David Kaplan [1989]: "Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and Other Indexicals". In Joseph Almog, John Perry and Howard Wettstein (Eds.) *Themes From Kaplan,* Oxford University Press, 481–563

Stefan Kaufmann [2009]: "Conditionals Right and Left: Probabilities for the Whole Family". *Journal of Philosophical Logic*, 38(1): 1–53

Niko Kolodny and John MacFarlane [2010]: "Ifs and Oughts". *The Journal of philosophy*, 107(3): 115–143

Christine M. Korsgaard [1983]: "Two Distinctions in Goodness". *The Philosophical Review*, 92(2): 169–195

David Krantz, Duncan Luce, Patrick Suppes and Amos Tversky [1971]: *Foundations of Measurement, Vol. I: Additive and Polynomial Representations*. New York Academic Press

David Lewis [1976]: "Probabilities of Conditionals and Conditional Probabilities". *The Philosophical Review*, 85: 297–315

— [1981]: "Causal Decision Theory". *Australasian Journal of Philosophy*, 59: 5–30

— [1988]: "Desire as Belief". *Mind*, 97: 323–332

— [1996]: "Desire as Belief II". *Mind*, 105: 303–313

Ted Lockhart [2000]: *Moral Uncertainty and Its Consequences*. Oxford University Press

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

William MacAskill, Owen Cotton-Barratt and Toby Ord [2020b]: "Statistical Normalization Methods in Interpersonal and Intertheoretic Comparisons". *Journal of Philosophy*, 117(2): 61–95

Matthew Mandelkern [2024]: *Bounded Meaning: The Dynamics of Interpretation*. Oxford University Press

Sarah Moss [2015]: "On the Semantics and Pragmatics of Epistemic Vocabulary". *Semantics and Pragmatics*, 8: 5:1–81

Graham Oddie [2001]: "Hume, the BAD Paradox, and Value Realism". *Philo*, 4(2): 109–122

Philip Pettit and Huw Price [1989]: "Bare Functional Desire". *Analysis*, 49(4): 162–169

Huw Price [1989]: "Defending Desire-as-Belief". *Mind*, 98: 119

Deniz Rudin [2025]: "Asserting Epistemic Modals". *Linguistics and Philosophy*, 48(1): 43–88

Jeffrey Sanford Russell and John Hawthorne [2016]: "General Dynamic Triviality Theorems". *The Philosophical Review*, 125(3): 307–339

Brian Skyrms [1982]: "Causal Decision Theory". *The Journal of Philosophy*, 79(11): 695–711

Michael Smith [1994]: *The Moral Problem*. Oxford: Blackwell

Robert Stalnaker [1975]: "Indicative Conditionals". *Philosophia*, 5(3): 269–286

— [2019]: "Expressivism and Propositions". In *Knowledge and Conditionals: Essays on the Structure of Inquiry,* Oxford: Oxford University Press

H. Orri Stefánsson [2014]: "Desires, Beliefs and Conditional Desirability". *Synthese*, 191(16): 4019–4035

Bas van Fraassen [1976]: "Probabilities of Conditionals". In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science,*

Brian Weatherson [2019]: *Normative Externalism*. Oxford, New York: Oxford University Press

Ruth Weintraub [2007]: "Desire as Belief, Lewis Notwithstanding". *Analysis*, 67(2): 116–122

Seth Yalcin [2007]: "Epistemic Modals". *Mind*, 116(464): 983–1026

— [2011]: "Nonfactualism about Epistemic Modality". In Andy Egan and Brian Weatherson (Eds.) *Epistemic Modality,* Oxford University Press, 295–332