Intrinsic Desire as Belief

Wolfgang Schwarz

Draft, June 2025

1 Introduction

Normative beliefs – about what is good or bad, right or wrong, about reasons, etc. – are connected to desire and motivation. For the most part, people are motivated to make the world better, not worse. Much has been written about the status and strength of this connection. Is it a conceptual truth? A norm of rationality? A contingent fact about human psychology? And can we make any sense of it, if we understand normative beliefs as ordinary beliefs about a special subject matter? Does it support the idea that normative beliefs are really some kind of desires in disguise?

This paper is about a different question: what is the *form* of the connection between normative belief and desire? More specifically, how can we spell out the connection in the framework of Bayesian decision theory, arguably our best model of idealised belief, desire, and motivation? Before we ask how something might be explained, it is useful to get clear about what, exactly, it is that we want to explain. The task is surprisingly difficult.

Suppose you are wavering between two normative theories, giving some credence to each. Theory One recommends that you do A, Theory Two that you do B. We might expect that the more strongly you believe in the first theory, the more you are motivated to do A, and vice versa. More precisely, it is tempting to assume that your degree of desire for a given option might equal the average of the value assigned to the option by the two theories, weighted by your degree of belief in each theory. This is a popular view in the literature on moral uncertainty (see, e.g., [Lockhart 2000], [MacAskill and Ord 2020]). An analogous model has been suggested for the connection between motivation and beliefs about reasons (see, e.g., [Gregory 2021: ch.8][1]). But can it be true?

An obvious difficulty is that the proposal requires us to find a common numerical scale on which we can compare the verdicts of different normative theories. This is a serious problem, but I am cautiously optimistic that it can be solved – perhaps by the statistical normalization techniques discussed in [MacAskill et al. 2020].[2]

1 Gregory suggests that if you give any credence to having a reason for A of strength x_1, and to 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 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 the average of the "strengths" of the desires, weighted by your credence in the corresponding reason hypotheses.

2 For more on this problem, see [Lockhart 2000], [Sepielli 2009], [Hedden 2016], [Carr 2020], [Tarsney et al. 2024],

I want to focus on a different problem. It comes to light if we assume a decision theory like that of [Jeffrey 1965], in which desires don't pertain to unstructured "options" or "outcomes", but to elements of the same algebra over which beliefs are defined.[3] The problem was first raised (albeit in a somewhat opaque way) in [Lewis 1988a] and [Lewis 1996], where Lewis argued that a certain anti-Humean "Desire as Belief" thesis is incompatible with Jeffrey's decision theory. These days, Lewis's argument is widely thought to rest on an obvious mistake – although there is no agreement on what that mistake is.[4] But the problem he uncovered is real.

Lewis's problem is closely related to another problem, first raised in [Smith 1994] and re-emphasized in [Weatherson 2014] and [Weatherson 2019]: making desires sensitive to moral beliefs seems to require a model of agents who don't intrinsically care about honesty or justice or the welfare of others; their only (morally relevant) intrinsic desire is to bring about whatever is morally good. I agree with Smith and Weatherson that there is something deeply wrong with this kind of "moral fetishism".

Fortunately, both problems can be solved. They key is to distinguish between *overall* desire and *intrinsic* desire, and between *overall* goodness and *intrinsic* goodness. There is no simple connection between overall desire and beliefs about overall goodness. Indeed, it is doubtful that we can accommodate beliefs about overall goodness at all in Bayesian models of belief. But we *can* find a simple connection between intrinsic desire and beliefs about intrinsic goodness.

My positive proposal will have to wait until section 9. Until then, I will explain what I see as the problem uncovered by Lewis, and why common responses to his argument don't resolve it. Section 2 reviews Lewis's argument. In sections 3–5, I will try to bring out the underlying problem, which turns on the "information-sensitivity" of the concept of overall goodness. In sections 6–8, I will explain why some initially attractive lines of response don't provide a satisfactory way out. Let's begin with Lewis's argument.

2 Lewis on Desire As Belief

We want to model the connection between normative beliefs and desire. Following Lewis, I will 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. I assume that goodness pertains to (possibly centred) propositions, which may describe acts or intentions or other normatively relevant events.

Plausibly, goodness comes in degrees. Let '$g(A) = x$' express that A is good to degree x – on some reasonable scale that allows comparing the verdicts of different normative theories.

We can now formalize the belief-desire connection floated in the previous section. Let Cr be

among others.

3 See, e.g., [Lewis 1981], [Skyrms 1982], [Joyce 1999], and [Bradley 2017] for other theories of this kind. For advantages of this approach, see also [Skyrms 1990] and [Joyce 2000].

4 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]).

the credence function and V the desirability function of a rational agent. Our hypothesis was that the agent's degree of desire towards a proposition A might equal the credence-weighted average (in probability jargon, the "expectation") of A's degree of goodness. That is, for all rational credence functions Cr, desirability functions V, and propositions A,[5]

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

This is (a notational variant of) thesis (17) in [Lewis 1988b]. To keep the maths simple, Lewis mostly concentrates on the case where there are only two degrees of goodness: 0 and 1, say. Ex then reduces to $V(A) = \mathrm{Cr}(g(A) = 1)$. Writing $\mathring{A}$ (read 'A is good') for $g(A) = 1$, we get the "Desire As Belief" thesis that is Lewis's main target:

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

DAB 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 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 you may rationally prefer rescuing both to rescuing one (and rescuing one to rescuing neither). But Lewis assumes (also for simplicity) that you 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 distinguish only two levels of goodness. That said, the bizarre assumption really does simplify the formalities – which is why I'll often stick to it in what follows, dealing with the more realistic general case in footnotes.

Here, then, is Lewis's refutation of DAB, as presented in [Lewis 1996]. Take any instance of DAB. 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

$$\mathrm{Cr}^A(\mathring{A}) = \mathrm{Cr}(\mathring{A} \ / \ A). \tag{Conditionalization}$$

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

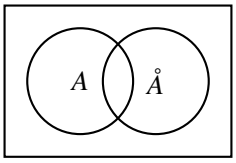
$$V^A(A) = V(A). \tag{Invariance}$$

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

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

5 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.

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



But picture A and $\mathring{A}$ as regions in logical space. Surely A and $\mathring{A}$ can often be true together. So take a case where the regions overlap, and where $\mathrm{Cr}(A) > 0$. Then $\mathrm{Cr}(\mathring{A}/A)$ is the ratio of the credence in the intersection $A \wedge \mathring{A}$ over the credence in A. Independence says that this ratio equals the total credence in $\mathring{A}$. This might be true for a particular credence function Cr, but if it is, we can easily make it false by moving around some of the credence. For example, we can increase $\mathrm{Cr}(\mathring{A})$ without changing $\mathrm{Cr}(\mathring{A}/A)$ by moving credence from outside $A \vee \mathring{A}$ into $\mathring{A} \wedge \neg A$, or by conditionalizing on $A \vee \mathring{A}$. 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 that 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 for special kinds of agents under "ideal conditions", provided the conditions are closed in the relevant way.[6]

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 (with a rather different style of argument) that even an ungraded formulation of DAB runs into trouble.

We've relied on the simplifying assumption that there are only two levels of goodness. This, too, is inessential. The refutation generalizes to Ex.[7]

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 Ex run into trouble. One is tempted to blindly tinker with the premises 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.

6 As [Lewis 1988a] and [Byrne and Hájek 1997] show, one can substantially weaken the closure conditions on Cr.

7 If Ex 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.

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* axiom in Jeffrey's decision theory. The axiom 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:[8]

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

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.

By DAB, $\mathrm{Cr}(\mathring{A})$ equals $V(A)$. This means that $\mathrm{Cr}(\mathring{A})$, too, must obey some kind of averaging principle: the credence that A is good must be sensitive to the distribution of credence over the subregions of A. This makes intuitive sense. If A has good parts and bad parts, the probability that A is good depends on the relative probability of these parts. If I'm confident that many lives will be saved if you donate to a certain charity, and you are confident that the charity is a scam that benefits no-one, we would expect that I'm more confident than you that donating to the charity is good.

Surprisingly, this intuitive thought can't be right: $\mathrm{Cr}(\mathring{A})$ can't be governed by the kind of averaging principle required by DAB or Ex. This is the problem that Lewis discovered.

The exact shape of the problem depends on what we say about V_b. $V(A)$ is determined by V_b and the distribution of credence over the subregions of A. By DAB, it follows that $\mathrm{Cr}(\mathring{A})$, too, is determined by V_b and the distribution of credence over the subregions of A. What can we assume about the basic desire function V_b?

The Lewisian problem arises for almost any way of answering this question. Nonetheless, it will be useful to have a concrete model on the table. I will suggest two models. The first assumes that the

See also [Lewis 1988a: sec.4].

8 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.

agent has a fixed, basic desire for goodness. The second assumes that the agent's basic desires are sensitive to their beliefs about goodness.

Here is the first model. 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. Setting aside the problem of intertheoretic value comparison, we can assume that a maximally specific normative proposition is represented as a value function v that assigns a goodness score to each world.[9]

Now consider a world in which, say, (a) the only ultimate good is that the number of stars is even, and (b) the number of stars is indeed even. If your only ultimate desire is that the norms – whatever they might be – are satisfied, then you should desire that this world is actual. It's a world where everything is as it ought to be, and that's what you want.[10] More generally, on this model you desire each world in proportion to the goodness level the world assigns to itself. Let v_w be the value function built into world w. The first model then fixes V_b by setting $V_b(w)$ equal to $v_w(w)$:[11]

$$V_b(w) = v_w(w). \tag{DBN+}$$

DBN+ is a fetishist model. To see what's wrong with it, let w be a world in which rape and murder are (a) pervasive and (b) good. Do you desire that this world is actual? I hope you don't. But $v_w(w)$ is high. By DBN+, $V(w)$ is high as well. The moral fetishist desires worlds like w. Good people, by contrast, don't just desire that the norms – whatever they might be – are satisfied. They have an intrinsic concern for the life and well-being of people. They abhor rape-and-murder worlds, no matter what these worlds say about their own value.

My second model takes this into account. Here, the agent uses their actual, unconditional beliefs about the norms to evaluate the desirability of a world. The simplest way to spell this out identifies the desirability of a world with its expected goodness. It's a form of Ex, but restricted to individual worlds:[12]

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

The model that I'll present in section 9 will entail Alignment. Our immediate task, however, is to get clearer about the refutation of DAB and Ex. For this, the choice between the two models makes no real difference.

4 Triviality

Let A be an "uneven" proposition with good and bad parts. We've seen that the desirability of A, as measured by $V(A)$, depends on the relative probability of these parts. Since DAB equates $\mathrm{Cr}(\mathring{A})$ with

9 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.

10 Compare [Lewis 1988a: p.332].

11 DBN+ is closely related to a principle that [Lewis 1996] calls 'Desire by Necessity'. See section 7 for discussion.

12 In Alignment, 'v' ranges over maximally specific normative propositions; I assume that each such proposition determines a goodness score for each world; I use '$v(w)$' for the score of world w under v.

$V(A)$, it implies that $\mathrm{Cr}(\mathring{A})$, too, depends on the relative probability of the parts of A. In fact, if we hold fixed V_b, $\mathrm{Cr}(\mathring{A})$ would have to be *determined* by the distribution of credence over the subregions of A. But how could this be, given that A and $\mathring{A}$ are logically independent?

Let's go though this more slowly. Assume that there are only two levels of goodness. DBN+ then reduces to

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

where G be the set of worlds w with $v_w(w) = 1$.[13] We also get an instance of DBN from Alignment if we assume that the agent is certain of the norms: in this case, G is the set of worlds that are good according to these norms.

Informally, G says that the world is good (either according to itself, or according to the norms of which the agent is certain). Together, DBN and Averaging entail

$$V(A) = \mathrm{Cr}(G/A).$$

This makes sense: if we want the world to be good, the desirability of a proposition A should be proportional to the probability that the world is good given A.

Now assume that $V(A) = \mathrm{Cr}(\mathring{A})$, as DAB demands. We then get an equality of belief with conditional belief:

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

This still looks plausible (modulo the implausible assumption of having only two levels of goodness). Again, if you think that donating to charity is good, you probably assign greater credence to the good subregions of that proposition than to the bad ones. If your credence shifts towards the bad subregions, you'll become less confident that donating to charity is good.[14]

But BACB can't hold in general. Like Independence, it may hold for a particular credence function, but it can then easily be rendered false by moving around credence. For example, 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

13 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 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 $V(w) = 0$.

14 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: 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.)

lies on the good part. By BACB, you are confident that A is good. If we move around your credence within A, so that most of it comes to lie on the bad part, BACB says that you 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 credence now lies outside $\mathring{A}$, while previously most of it lay inside. This is impossible!

Note that this argument doesn't involve any updating. I'm not comparing an earlier and a later credence function. I haven't assumed Invariance. Apart from DAB and Jeffrey's Averaging axiom, I've only assumed that we can identify a proposition G of which it is known that it comprises all and only the good worlds. DBN+ ensures that there is such a proposition. With Alignment, we had to consider a case where the agent is sure of the norms, but it's not hard to see that this assumption could be dropped: we can dispense with G.[15] We can also dispense with the assumption that there are only two levels of goodness: the argument works just as well with Ex in place of DAB.[16] There's nowhere to hide. Given the Averaging axiom for desire, DAB and Ex can't be true (unless we drastically restrict the scope of Cr and V).

The problem generalizes to other ways of connecting beliefs about goodness to desire. Consider this comparative version of DAB, where V_1, Cr_1 and V_2, Cr_2 are two pairs of a rational credence and a desirability function:

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

15 Obviously, it would be enough that the agent is sure what the norms say about the worlds in A. But we don't even need that. Alignment and Averaging together entail that

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

By DAB, this yields

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

Now $\mathrm{Cr}(\mathring{A})$ doesn't just depend on the distribution of credence within A. Outside A, however, $\mathrm{Cr}(\mathring{A})$ only depends on the credence distribution over normative matters. And that's not enough to avoid the refutation by redistributing credence.

16 With more than two levels of goodness, DBN+ and Averaging yield a version of the "Desire As Expected Goodness" thesis defended in [Broome 1991]:

$$V(A) = \sum_x x \, \mathrm{Cr}(G_x/A),$$

where G_x is the set of worlds with self-assigned goodness level x. If we combined this with Ex, we get another equation between an unconditional and a conditional expectation:

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

This leads to the same kind of trouble as BACB: the right-hand side only depends on the relative distribution of credence inside A, while the left-hand side also constrains the distribution outside A. We can therefore break the equality by (for instance) moving around credence outside A so as to change the left-hand side, without moving around credence inside A. Essentially the same maneuver works with Alignment instead of DBN+. In that case, Ex yields

$$\sum_x \mathrm{Cr}(g(A) = x)\, x = \sum_w \left(\mathrm{Cr}(w/A) \sum_v \mathrm{Cr}(v)v(w) \right).$$

This, too, only holds for an excedingly narrow range of credence functions.

Focus again on a case where DBN holds (either because we assume DBN+ or because the agent is sure about the norms). As above, we then have $V(A) = \mathrm{Cr}(G/A)$. Comparatie DAB then turns into:

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

Again, this looks plausible (modulo the two levels of goodness). If I am more confident that good things happen if you donate to charity than you, we'd expect that I am more confident that donating to charity is good. But it can't hold in general. For example, assume that $\mathrm{Cr}_1(G/A) > \mathrm{Cr}_2(G/A)$, and that Cr_1 and Cr_2 both give low credence to 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 $\neg\mathring{A}$.[17]

We can also consider direct comparative judgements. Let '$A \succ B$' express 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:[18]

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

From $V(A) = \mathrm{Cr}(G/A)$, we could infer that

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

Once more, this looks plausible: if you're sure that A is better than B you'll think it's more likely that good things happen given A than given B. And once more, it can't hold in general. If $A \succ B$ overlaps G, we can falsify it by concentrating Cr on $(A \succ B) \wedge G$; we'll then have $\mathrm{Cr}(A \succ B) = 1$, but $\mathrm{Cr}(G/A) = \mathrm{Cr}(G/B) = 1$. If $A \succ B$ doesn't overlap G, we can falsify it by concentrating Cr on $A \succ B$: we'll have $\mathrm{Cr}(A \succ B) = 1$ and $\mathrm{Cr}(G/A) = \mathrm{Cr}(G/B) = 0$.

These are "triviality arguments". They don't strictly refute the suggested principles. But they show that the principles can only hold unrestrictedly under trivial conditions – for example, if all worlds are equally good.

I want to stress how perplexing these results are. They don't just refute a dubious combination of contentious hypotheses from metaethics and decision theory. They refute assumptions that may have seemed obviously true – or at least obviously coherent.

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

17 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 $\mathrm{Cr}_1(\mathring{A}) < \mathrm{Cr}_2(\mathring{A})$.

18 [Russell and Hawthorne 2016: 315f.] consider a closely related principle ("Comparative Value"). Using a dynamic argument, they show that it leads to $\mathrm{Cr}(A \wedge B \wedge (A \succ B)) = 0$.

evaluate the probability of B on the supposition that A. But Stalnaker's Thesis fails for the same reason as BACB (which is a special case of the 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 construed for 'must' and 'might', on their epistemic reading.[19] All these cases – conditionals, epistemic modals, statements about goodness – have something in common. Intuitively, the relevant statements make a claim (in part) 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. 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.[20] That's what gives rise to the trouble.

The exact mechanisms of how all this works are still a matter of active research. As far as I can tell, there is, however, a near consensus that one can't give an adequate account of information-sensitive expressions within the strict 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.[21]

Our topic isn't the semantics of information-sensitive expressions. We want to know how beliefs about goodness relate to desire and motivation, in a decision-theoretic framework. It's a presupposition of our project that the relevant beliefs about goodness can be modelled as attitudes (credences) towards sets of worlds.[22] The upshot, I think, is that we need to use a conception of goodness that isn't information-sensitive – a conception on which the goodness of a proposition does not depend on the relative probability of its parts. In section 9, I'll argue that we do have such a conception: *intrinsic goodness*. Before we get there, I want to explain why some other escape routes are dead ends.

5 Escape from triviality?

Let's go through the refutation of DAB from the previous section once more, to see how it might be resisted. We assume DBN, so that $V(\mathring{A}) = \mathrm{Cr}(G/A)$. Assume also that A divides into a G part and a $\neg G$ part, and that the overall probability of A is low. According to DAB, 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, the probability of $\mathring{A}$ is low. If we move the probability in A onto the G part, the probability of $\mathring{A}$ becomes

19 See, e.g., [Russell and Hawthorne 2016], [Goldstein and Santorio 2021].

20 The linguistic environment can further manipulate the information parameter; we'll see examples in the next section.

21 See, among many others, [Yalcin 2011], [Moss 2015], [Stalnaker 2019], [Hawke and Steinert-Threlkeld 2021], [Mandelkern 2024], [Ciardelli and Ommundsen 2024], [Rudin 2025]. 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]).

22 When I speak of "worlds", I always mean atoms in the algebra over which credence is defined (assuming, for simplicity, that the algebra is atomic). It doesn't matter if these atoms are metaphysically possible.

high. But moving a small amount of probability inside A can't change any unconditional probability from low to high.

One way to block this argument is to simply forbid any movement of credence that would render DAB 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 block the argument drops the assumption that $\mathring{A}$ is a fixed proposition. What if $\mathring{A}$ is a proposition with low probability if most of the probability in A lies on the $\neg G$ part, but it becomes a different proposition, with high probability, if we move that probability onto G?

Both of these ideas have been explored in the literature on conditionals, in response to the triviality arguments against Stalnaker's Thesis.[23] They face numerous problems.[24] 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. It'll be worth explaining this briefly.

Consider the following setup.[25] 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, the probability is 1: $\mathrm{Cr}(\mathit{Low} \Rightarrow \mathit{Two}/\mathit{Even}) = 1$. By probability theory, 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.

We could reject the intuitive judgement, but that judgement does not stand alone. Consider, for example, 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.[26] Or consider the probability of $\mathit{Even} \wedge (\mathit{Low} \Rightarrow \mathit{Two})$. Intuitively, this equals the probability of *Even*: if the outcome is even then it must be two if it is below four, so $\mathit{Low} \Rightarrow \mathit{Two}$ doesn't add anything to *Even*. 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 inconsistent 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

23 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].

24 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.)

25 This illustration is due to [Ciardelli and Ommundsen 2024], who closely follow [Goldstein and Santorio 2021], who in turn follow [Fitelson 2015].

26 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 seem at least as plausible as Stalnaker's Thesis.

and $\mathrm{Cr}(A \Rightarrow B/D)$ low, even though the probability of $B \vee D$ is moderately high.[27] These judgements, too, are probabilistically inconsistent with Stalnaker's Thesis.

Unlike Lewis's triviality argument, these arguments don't apply Stalnaker's Thesis to an updated credence function. They don't involve any movement of credence at all. As a consequence, neither of the two responses to the triviality results can escape them. Note also that the relevant instances of Stalnaker's Thesis are intuitively *true* in these cases. The probability of *Low* ⇒ *Two*, for example, really seems to be 1/3. Even for a single, fixed probability function, there is no proposition whose probability could match our intuitive judgements about the probability of a conditional $A \Rightarrow B$.

We can construct similar static arguments to show that our intuitive judgements about goodness are probabilistically incoherent. 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 won't keep rising for the next 50 years *or* greenhouse gases have a net positive effect on the environment. Suppose we think it's likely that greenhouse gases won't keep rising, but highly unlikely that they are net positive. So $\mathrm{Cr}(B)$ is high. $\mathrm{Cr}(\mathring{A})$ is low: it's not good if greenhouse gases keep rising. (This is in line with DAB.[28]) On the other hand, $\mathrm{Cr}(\mathring{A}/B)$ is intuitively high: if B is true then runaway greenhouse emissions are sure to have good consequences. But probability theory requires that $\mathrm{Cr}(\mathring{A}/B) \leq \mathrm{Cr}(\mathring{A}) / \mathrm{Cr}(B)$.[29]

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 will keep rising but are net positive. The hypothesis is unlikely, but not completely impossible: $\mathrm{Cr}(A \wedge B) > 0$. 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.[30]

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. They are either in shaft A or in shaft B; we don't know which. We can block at most one shafts from being flooded. Without further information, 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

27 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].

28 Take a simple case where we have DBN, so that DAB entails that $\mathrm{Cr}(\mathring{A}) = \mathrm{Cr}(G/A)$. Plausibly, $\mathrm{Cr}(G/A)$ is low, given an obvious way of filling in Cr.

29 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.

30 [Weintraub 2007] skirts this puzzle. By Averaging, $V(A) = V(B)$ whenever $\mathrm{Cr}(A \wedge B) = 1$. 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.].

probability theory: if 'N is best' has high unconditional probability, it can't have low probability conditional on each of two hypotheses whose disjunction has high probability.

These cases show that even for a single, fixed probability function Cr, our judgements about the probability that A is good (or better than B, or best) can't be understood as judgements about the probability of some proposition expressed by 'A is good' (etc.). The cases also illustrate an important fact about information-sensitive expressions: when we assess the probability of an information-sensitive statement, the linguistic environment can manipulate the information 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.[31]

At this point, I hope you agree that the prospects for upholding anything like Ex or DAB look dim. We need to rethink the connection between normative belief and desire. I'll discuss two such proposals in sections 7 and 8, before turning to my own proposal. First, though, I want to quickly comment on the role of the Invariance assumption in Lewis's argument, which has sometimes been singled out as the culprit.

6 Invariance

My argument against DAB relied on Jeffrey's Averaging axiom, by which the desirability of a proposition depends on the distribution of credence over its parts. More precisely, if we hold fixed the basic desire function V_b, Averaging entails that $V(A)$ is determined by the distribution of credence within A. But $\mathrm{Cr}(\mathring{A})$ can hardly be so determined, as we saw when we looked at a case where $\mathrm{Cr}(A)$ is low. So $\mathrm{Cr}(\mathring{A})$ can't equal $V(A)$.

In his articles on Desire as Belief, Lewis assumes Jeffrey's theory, but his main argument does not make use of Averaging. The only assumption about V used in Lewis's argument 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 together with the assumption that conditionalizing on some information leaves an agent's basic desires V_b unchanged:[32]

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

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

31 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.

32 Suppose we conditionalize the credence function Cr on some hypothesis E, without changing V_b:

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

especially undesirably parts of a proposition. Obviously, this move would also make no real difference to my argument against DAB: even with risk-weighting, $V(A)$ will only depend on the distribution of credence within A. And this much is entailed by Invariance: if $V(A)$ equals $V^A(A)$, it can't depend on the credence outside A, as that part of Cr leaves no trace after conditioning on A.[33]

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 that we could use 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. If we accept Normalization, we must therefore 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 this issue can be found in [Jeffrey 1977]. Jeffrey argues that we face a choice between two initially attractive assumptions: Stability and Normalization. (Remember that Stability entails Invariance.) He also argues that the choice is a philosophically insignificant matter of scaling. If we accept Stability, holding fixed the basic desire function V_b when the agent acquires information, we measure the new desirabilities on the same scale as the old desirabilities. If instead we hold fixed the neutrality of the tautology, we measure the new desirabilities on a different scale in which the zero is moved to the previous desirability of the information that has been acquired.

As one might expect, the choice makes no substantive difference to the refutation of DAB. My argument against DAB made no use of Invariance. 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.

7 Desire as Conditional Belief

[Price 1989] offers an attractive suggestion for how to rethink the connection between normative

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

33 Technically, Invariance allows that $V(A)$ doesn't depend on the distribution of credence within A either, but then it

belief and desire. Price points out that the calculation of V involves conditional credence, while DAB links V to unconditional judgements about goodness. The region outside A therefore plays a role for the assessments of A's goodness, but not for $V(A)$. This is what led to the trouble. To avoid it, Price suggests that we should equate $V(A)$ with the conditional credence that A is good, given that A is true:[34]

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

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, especially given that, a few pages earlier, he described 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: it follows from DBN+. In this case, G comprises all worlds that are good according to themselves. Lewis argued that friends of DAB should accept DBN+ ([Lewis 1988a: p.332]). He also pointed out that DBN is entailed by DAB ([Lewis 1996: sec.6]). So he could hardly have thought that it would be problematic if friends of DAB had to accept DBN. His objection, I think, is that they need to say *more* than DBN.

Early in [Lewis 1996], Lewis distinguishes two kinds of anti-Humeanism. One rejects Hume's *liberalism about desire*: his view that 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 on the *independence of belief and desire*: 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 mentions that he finds the proposed constraint on basic desire implausible, at least if understood as an analytic truth. But the proposal is obviously compatible with decision theory. That's the sense in which DBN is "simple and unproblematic".

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.

This argument against DACB can be strengthened. Remember that 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 express a constant function, mapping every proposition A to

wouldn't depend on Cr at all, which is incompatible with any DAB-type link.

34 For more than two levels of goodness, we would say that $V(A) = \sum_x \mathrm{Cr}(g(A) = x/A)$.

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 $\mathring{A}$ with G. Given DBN, however, it requires identifying $\mathring{A}$ with G *inside* 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 DBN: what I called his "complicated proof" shows that if DACB holds unrestrictedly then there must be a fixed proposition G such that for all A, $\mathring{A} \wedge A = G \wedge A$.[35] 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.[36] But that's not true. We know more about goodness (and other normative properties). We 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 $\mathrm{Cr}(A \wedge B / \neg\mathring{A} \wedge \mathring{B}) = 0$.[37] 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

35 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 $\mathring{A} \wedge A = \mathring{\top} \wedge A$.

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

37 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 $A \wedge B$.

DACB, just as we know more about indicative conditionals than what's entailed by Stalnaker's Thesis. What we know about goodness seems to contradict DACB.

For an even clearer example, return to the rape-and-murder world w that regards itself as good. I claim that w is not good; nor is it desirable. This directly contradicts DACB, for w is certainly good conditional on itself: $\mathrm{Cr}(\mathring{w}/w) = 1$.

We have yet to find a way out.

8 Objective value, subjunctive desire

Averaging says that when we consider the desirability of a proposition, we should weight its 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

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

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*: $\mathrm{Cr}(\cdot//B)$ is the probability that results when moving the probability mass of any $\neg B$-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:

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

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:[38]

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

38 For more than two levels of goodness, we would have $U(A) = \sum_x x\,\mathrm{Cr}(g(A) = x)$.

Does this share the fate of DAB? 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. Now suppose we accept DBN+, so that $V_b(w) = 1$ if $v_w(w) = 1$, otherwise $V_b(w) = 0$. Let $G = \{w \mid v_w(w) = 1\}$, as above. It follows that

$$U(A) = \mathrm{Cr}(A//G) = \mathrm{Cr}(A > G).$$

If we now define $\mathring{A}$ as $A > G$, we get SDAB![39]

We can escape the triviality problems because $\mathring{A}$ has become information-independent. $A > G$ 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 have there reservations.

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 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 $\mathrm{Cr}(\mathring{A}) \neq \mathrm{Cr}(G//A)$, no matter how $\mathring{A}$ is defined. Since $U(A) = \mathrm{Cr}(G//A)$, SDAB fails.[40]

Second. The above construction assumes that we can 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 SDAB 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 $A > G$.

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

40 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 $\mathrm{Cr}(G//A) = 0.5$.

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.

Above we assumed DBN+. But DBN+ is fetishist. It gets the rape-and-murder case wrong. Suppose we switch to Alignment. Combining SDAB with Alignment, we get

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

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 any region in logical space, we must have $w \in \mathring{A}$ iff $\mathrm{Cr}(\mathring{A}/w) = 1$, provided $\mathrm{Cr}(w) > 0$. By SBACB*, $\mathrm{Cr}(\mathring{A}/w) = 1$ 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, $\mathring{A}$ and S must coincide. Now A and $\mathring{A}$ shouldn't always be inconsistent with one another, so there should be cases where S contains some A-world. 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 SBACB*, $\mathrm{Cr}(\mathring{A})$ 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 SDAB with independently 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. 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. The above defence of SDAB doesn't deliver this. It assumes that all agents have a basic desire for G. A belief that $A > G$ is then 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?[41] 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

41 The following explication is closer to what [Korsgaard 1983] calls 'final goodness' than what she calls 'intrinsic good-

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 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.[42] 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.[43] (We can allow for agents with extreme bouletic holism who intrinsically care only about a single, very

ness'. 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.

42 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.

43 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.

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.

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:[44]

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

We might now posit an "intrinsic" form of Ex, identifying the intrinsic desirability of a proposition with its expected intrinsic goodness:

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

IEx implies that if you're sure that A is intrinsically good to degree x then you 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, IEx reduces to

$$V_i(A) = \mathrm{Cr}(\overset{\star}{A}), \qquad \text{(IDAB)}$$

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

My formulation of IEx assumes that all candidate value functions v assign an intrinsic goodness score to A. This need not be the case. To make the equation well-defined, 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. (So $V_i(A)$ 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.)

IEx is closely related to the Alignment principle for V_b:

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

In particular, Alignment follows from IEx together with the Additivity principles, for V and for Good-

44 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. As in section 6, I assume that v_i ranges over the propositions in certain families, each of which represents a

ness.[45]

This clarifies how a model based on Alignment avoids the charge of fetishism. "Good people", writes [Smith 1994: p.75], "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". In our model, these non-derivative concerns are represented as intrinsic desires. An agent who conforms to IEx cares non-derivatively about honesty iff they believe that honesty is intrinsically good. More precisely, the strength with which they care non-derivatively about honesty is proportional to their expectation of honesty's intrinsic goodness.

This may seem strange. If your concern for honesty depends on your beliefs, isn't it derivative? It depends on the nature of the dependence. Your concern for honesty plausibly depends on the presence of oxygen in your environment: without oxygen, you would soon have no concerns at all. But this doesn't make your concern derivative. The dependence expressed by IEx might be of the same kind.[46] It certainly isn't the kind of dependence expressed by DBN+, where your concern for honesty would be rationally explained by an underlying concern for goodness. DBN+ and Alignment make demonstrably different predictions.

We have to be careful when we think of V as a measure of "news value". Consider once more the hypothesis H that rape and murder are (a) pervasive and (b) intrinsically good. An agent who

(separable) dimension of value. We can allow for extreme value holism, where there is only a single, fine-grained family.

45 By Additivity,

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

Combining this with IEx, 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,

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

So

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

Conversely, we can almost derive IEx 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 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 IEx:

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

I say that this is *almost* a derivation of IEx 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.

46 IEx is just an equality. It doesn't say *why* the two sides are equal.

conforms to IEx may well find H abhorrent: $V(H)$ might be low. But if they were to learn that H is true, and if they still conformed to IEx, they would change their intrinsic desires. $V(H)$ would become high. As a mere possibility, H is "bad news". But the information that H is true 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. Here, too, we have to be careful. It's clear what Cr^E is supposed to be, but what is 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, if the agent conforms to IEx both before and after the update. On another – equally important – 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.[47] On this reading, Invariance and Stability hold, but we can't assume that the V^E, Cr^E pair satisfies IEx, for V^E doesn't process the information in E to revise the intrinsic desirability function, as would be required by Ex if the agent actually learned E.

10 Overall desire, overall goodness

If we combine Alignment with Averaging, we get

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

where v ranges over maximally specific hypotheses about goodness, and $v(w)$ is the goodness of w according to v. This is the overall (non-intrinsic) desirability of a proposition A.

In section 6, I mentioned that switching from DBN+ to Alignment does not help with the triviality problem: we still run into triviality if we set $\mathrm{Cr}(\mathring{A})$ equal to $V(A)$. The concept of goodness expressed by the halo would have to be information-sensitive.

But we *do* have an information-sensitive concept of goodness. Let's briefly explore how it might work.

On the model from the previous section, purely normative hypotheses don't directly assign an "overall goodness" score to arbitrary propositions. Rather, they assign an intrinsic goodness score to certain aspects of worlds, from which we can derive (by Additivity of Goodness) a value $v(w)$ for individual worlds w. How could we extend such a value function v so that it gives an "overall" verdict about arbitrary propositions, which typically divide into better and worse parts? To give a sensible verdict, v must be supplied with a probability measure over these parts. Let v^{Cr} be the value function induced by v and Cr so that

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

47 See [Bradley 1999], [Bradley 2017: ch.6], [Joyce 1999: ch.6], and [Joyce 2020] for more on conditional desirability

This equals $V(A)$ for an agent with credence Cr who is sure that v is the true value function.

We can now define a simple information-sensitive concept of goodness, g^{Cr}, by stipulating that at any world w with built-in value function v_w, $g^{\mathrm{Cr}}(A) = v_w^{\mathrm{Cr}}(A)$. That is, '$g(A) = x$' is true at a world, relative to Cr, iff the value function built into that world, extended by Cr, assigns x to A. '$g(A) = x$' expresses a proposition (a set of worlds) only relative to a credence function.

Now remember that when we assess the probability of simple information-sensitive statements A, we seem to plug the 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 "overall" good to degree x, by assuming that we evaluate $\mathrm{Cr}(g^{\mathrm{Cr}}(A) = x)$.

To illustrate how this works, 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 1 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, represented by 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 $v_1^{\mathrm{Cr}}(\textit{Flip})$ and $v_2^{\mathrm{Cr}}(\textit{Flip})$. Let's say that $v_1^{\mathrm{Cr}}(\textit{Flip}) = x_1$ and $v_2^{\mathrm{Cr}}(\textit{Flip}) = x_2$. Intuitively, x_1 is the goodness of *Flip* according to v_1. Indeed, $\mathrm{Cr}(g^{\mathrm{Cr}}(\textit{Flip}) = x_1/v_1) = 1$. Likewise, $\mathrm{Cr}(g^{\mathrm{Cr}}(\textit{Flip}) = x_2/v_2) = 1$. The *expected* goodness of *Flip*, in our state of indecision between v_1 and v_2, is

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

We might think that this should match $V(\textit{Flip})$. 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

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

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.

In the example, this means that

$$V(Flip) = x_1 Cr(v_1) + x_2 Cr(v_2).$$

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 there is no such thing as $g(A)$; there is only $g^{\mathrm{Cr}}(A)$.

The present toy model obviously does not amount to a worked-out semantics for our information-sensitive concept of goodness. But I hope it explains how such a concept might work. In particular, it explains why Ex, despite its severe formal problems, looks so innocent.

11 Conclusion

It is a Moorean fact that we have normative beliefs, and that these beliefs are linked to desire and motivation. What, exactly, does this connection look like?

This is the question we have studied. It is harder than one might have thought. Since most propositions are "uneven", dividing into better and worse parts, a measure of goodness for such propositions arguably requires information about the likelihood of the parts. Beliefs about goodness thereby become entangled with other beliefs, in a similar way in which beliefs about conditionals are entangled with beliefs about antecedent and consequent, and this throws a wrench into a precise model of these beliefs. This is the problem that lies behind Lewis's triviality results.

We can bypass the problem by focusing on individual worlds. On the most straightforward model, each world has a fixed, built-in goodness level (given, for example, by $v_w(w)$, where v_w is a value function determined by w), and agents have a basic desire for worlds with high goodness levels. Such agents will desire a non-world proposition in proportion to their expectation of the goodness of the world, conditional on that proposition. We might even find a simple, direct link between an agent's subjunctive desires and their beliefs about how good the world would be if the relevant proposition were true.

But this kind of model is unattractively fetishist, as highlighted by the rape-and-murder case. Good people don't desire worlds where appalling things happen, even if the appalling things are good according to those worlds.

I've suggested a different kind of model, linking intrinsic desires with beliefs about intrinsic goodness. I've described a simple instance of this kind, where intrinsic desire equals expected intrinsic goodness. More complicated (and more realistic) variations are easy to construct, by adjusting principle IEx.

That principle is a genuine addition to decision theory. On reflection, I think it should have been clear from the outset that we should seek some such addition. Remember that V and U are determined by V_b and Cr. If we fix V_b, there is no degree of freedom left: any connection between Cr on the one hand and V or U on the other is either entailed by the Averaging rules or incompatible with these rules. If we want a model in which normative beliefs are non-trivially and non-redundantly linked to desire, we need a link between V_b and Cr.

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

Jennifer Rose Carr [2020]: "Normative Uncertainty without Theories". *Australasian Journal of Philosophy*, 98(4): 747–762

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 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

Brian Hedden [2016]: "Does MITE Make Right? On Decision-Making under Normative Uncertainty". In Russ Shafer-Landau (Ed.) *Oxford Studies in Metaethics, Volume 11,* chapter 5. Oxford: Oxford University Press, 102–128

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

— [2000]: "Why We Still Need the Logic of Decision"

— [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

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

— [1988b]: “Desire as Conditional Belief”

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

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

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

William MacAskill and Toby Ord [2020]: “Why Maximize Expected Choice-Worthiness?” *Noûs*, 54(2): 327–353

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

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

Andrew Christopher Sepielli [2009]: “What to Do When You Don’t Know What to Do”. In Russ Shafer-Landau (Ed.) *Oxford Studies in Metaethics, Volume 4,* chapter 1. Oxford: Oxford University Press, 5–28

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

— [1990]: *The Dynamics of Rational Deliberation*. Cambridge (Mass.): Harvard University Press

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

Christian Tarsney, Teruji Thomas and William MacAskill [2024]: "Moral Decision-Making Under Uncertainty". In Edward N. Zalta and Uri Nodelman (Eds.) *The Stanford Encyclopedia of Philosophy,* Metaphysics Research Lab, Stanford University, spring 2024 edition

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

Brian Weatherson [2014]: "Running Risks Morally". *Philosophical Studies*, 167(1): 141–163

— [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