A Lewisian Defence of Desire as Belief

Abstract. David Lewis (1988a, 1996) rejected the *Desire as Belief* thesis because it either imposes implausible normative constraints on desires or else requires evaluative beliefs to be unresponsive to evidence. However, we contend that Lewis' dispositional theory of value (1989) commits him to a restricted version of the principle, which we call *Euthymia*: rationality requires that our beliefs about the good correspond to our second-order desires. We argue that Euthymia avoids Lewis' objection and vindicates the possibility of improving evaluative outlooks through rational revision. The upshot is a refined model of rational desire that reconciles a broadly Humean view of motivation with normative appraisal at the second-order level.

1 Introduction

Is there a normative connection between what we desire and what we believe to be good? According to the *Desire as Belief* thesis, there is: desiring A involves believing that A is good. David Lewis famously rejected this view, arguing that it leads to contradiction when combined with standard principles of rational belief updating. As a result, Lewis defended a broadly Humean theory of motivation according to which beliefs and desires play fundamentally distinct roles in practical reasoning. However, we think that Lewis was too quick to reject Desire as Belief altogether. Indeed, given Lewis' (1989) own dispositional theory of value—in addition to the general advantages of connecting evaluative beliefs and desires—a restricted version of Desire as Belief should be attractive on Lewisian grounds.

One motivation for connecting desires with beliefs is that such a connection might serve to illuminate structural requirements on belief–desire combinations.

Certain failures of coherence between belief and desire resemble familiar forms of *akrasia*. Traditionally, akrasia concerns practical rationality and refers to a mismatch between one's evaluative judgment and intention to act. More recently, philosophers have discussed the analogous phenomenon in belief, epistemic akrasia, where one's first-order belief conflicts with one's higher-order belief about what one ought to believe.

While contested, many have argued that akrasia is irrational. To believe that one ought to φ while failing to intend to φ, or to believe that A while also believing that one should not believe that A, is, in one way or another, a normative shortcoming. In the same way, Desire as Belief may be a rational requirement: it seems irrational to believe that A is good while failing to desire it; or to desire that A while believing it is not good.

Still, one might resist this idea—as Lewis explicitly did—and insist that any belief–desire pair is rationally permissible. Your desire for an even number of blades of grass in your garden is simply not constrained by your belief about the goodness of an even number. Having odd desires is just that, odd; it is not a normative failure of any kind. However, not all desires are equal. As we argue below, Lewis' metaethical views commit him to holding that at least some desires are normatively tied to evaluative beliefs.

In metaethics, Lewis (1989) defended what he in one place called 'subjectivism with bells and whistles' (1993). His theory is subjectivist since it explains values in terms of the dispositions of those who value. The bells and whistles concern the idealisations Lewis introduces to make room for modesty: only the immodest take their current values to determine what is valuable. Roughly, what is valuable is that which a certain kind of idealised version of yourself desires to desire. However, it turns out that this metaethical view is in tension with Lewis' wholesale rejection of Desire as Belief.

In this paper, we first argue that Lewis' metaethical outlook commits him to a

principle which links second-order desires and evaluative beliefs. By analogy with the Enkratic principle—that first-order beliefs are rationally required to cohere with higher-order beliefs about what is rational—we call the resulting principle *Euthymia*:[1] your beliefs and the good should cohere with your second-order desires.

Secondly, we argue that Euthymia can escape Lewis' own objection to Desire as Belief. In particular, using a notion of stability, popular among metaethical expressivists, we argue that Lewis' dispositional theory of value entails that our second-order desires should aim to be stable.[2] Given such a regulative ideal, both evaluative beliefs and second-order desires should be evidence responsive. Lewis' counterexample to Desire as Belief is thereby avoided.

Our aim is not merely interpretive. We use Lewisian ideas to develop a new account of the relationship between value, motivation, and belief that resists his own objection while preserving the possibility of evaluative improvement. Specifically, we argue that while ideal and non-ideal agents alike should aim at such stability, rationality requires them to remain open to changing both their evaluative beliefs and second-order desires in response to relevant evidence.

As a result, Euthymia functions as a normative ideal which rational agents aim for through revision to their mental states. This vindicates Lewis' cognitivism about value: our evaluative outlooks can improve through evidence responsiveness. The upshot is that Lewis' objection is not only avoidable for second-order desires—it *is* possible to link a subset of desires and beliefs—but that this should indeed be welcome news for Lewis himself and for anyone who connects value with forms of epistemic idealisations.

---

[1] Etymological gloss: where *akrasia* literally means 'lack of mastery' and *enkratic* 'good mastery', *euthymia* literally means 'good passions'.

[2] Below we outline how our proposal aligns with expressivist views that appeal to epistemic idealisations to explain concepts like fallibility (e.g., Bex-Priestley 2018; Gamester 2022). But such appeals are not exclusive to expressivists. Naturalist realists like Railton (2017) may also find the view attractive.

In §2, we outline the objection Lewis raises to Desire as Belief. We reconstruct Lewis' dispositional theory of value in §3 and then, in §4, argue that it commits him to a restricted version of Desire as Belief which escapes his objection. In §5 we compare our solution to two proposals found in the literature. We briefly conclude in §6.

2 Lewis' argument against desire as belief

Sometimes we act contrary to what we desire. I desire ice cream, yet refrain. You desire to sleep in, yet get up to go to work. At first blush, this seems to violate the Humean theory of motivation according to which only desires motivate action.

The Humean response is obvious: perhaps there are 'underlying' desires for being healthy or doing one's job which motivate our actions. As Lewis (1988a) points out, Humeans are free to define desire so broadly as to include the underlying mental states that make us refrain despite temptation or get up despite the cold. If so, these are merely cases of conflicting desires. I desire both ice cream and health, and one desire wins out in motivating me to act. No mystery here for the Humean.

However, there is an alternative way of describing such cases which challenges the Humean analysis. Perhaps these cases involve *beliefs* about what is good, which bring about corresponding *desires* to achieve that good. Perhaps it is my belief that health is better for me which brings about my desire to stay healthy, thereby motivating me to refrain.[3]

The disagreement between Humean and anti-Humean views is unlikely to be settled by appeal to everyday examples about diets or sleep routines. More revealing are cases involving apparent conflict between our beliefs about the good and our

[3] While Lewis does not cite specific articulations of this view, the theory that desires are beliefs about what is valuable or good—so-called good-based theories of desire—is among the most prominent in the contemporary landscape. See Schroeder (2020) for a taxonomy and Gregory (2021) for a defence.

desires. Consider Augustine who steals pears even though he believes it is sinful. Or Kant's shopkeeper, who desires to cheat his customers but refrains solely because he knows it would be wrong. A cost for the Humean view is that it treats Augustine as perfectly rational (even if morally culpable) and it treats cases which the Kantian sees as exemplars of moral virtue as psychologically abnormal, if not downright metaphysically impossible (cf. Sinhababu 2017).

The anti-Humean need not deny the central role desires play in decision-making. When we are motivated by beliefs about the good, we are so indirectly: our beliefs about the good constrain what desires we have. Our beliefs about the good can then be the ultimate source of our motivation (see Gregory 2021: Ch. 3). Since Humeans claim that the source of our motivations is ultimately our desires, this conflicts with their view—at least if we only consider *non-instrumental* desires and beliefs about non-instrumental goodness. Humeans and anti-Humeans typically agree that our instrumental desires are at least to some extent rationally constrained, so in order to avoid obfuscating the debate, we will only focus on non-instrumental desires and beliefs about the non-instrumentally good.

One way to be anti-Humean then, is to maintain that, at least sometimes when we are motivated, our desire should be isomorphic to some of our beliefs about what is good. We can characterise such beliefs by their evaluative contents. Following Lewis (1988a), we can denote evaluative propositions with a propositional letter to which we add a 'halo'. The proposition $\mathring{A}$ is then the proposition expressing that it is or would be non-instrumentally good that A—good *simpliciter* that A. The degree to which you believe that it would be good that A is represented by the credence you give $\mathring{A}$.

We can now formulate the anti-Humean Desire as Belief thesis which Lewis seeks to refute. Where $\mathcal{V}$ is the desirability function of a given agent, which assigns a real number to each proposition, it represents their desire that the proposition obtains, and $\mathcal{C}$ is the credence function of the same agent, Desire as Belief is formulated by

Lewis (1996) as:

DAB: $\mathcal{V}(A) = \mathcal{C}(\mathring{A})$.

This anti-Humean view, then, says that one's desire for A equals one's credence that A is good.[4] Two qualifications are in order before we proceed to Lewis' objection to DAB.

What is the ontological status of the equality? Is the mental state of desiring A the very same mental state as the degree of belief assigned to $\mathring{A}$? Or is there rather some kind of necessary connection between these separate mental states? Given Lewis' general rejection of necessary connections between distinct existences (see Lewis 1986: §1.8), and his commitment to interpretationism about mental states (see Lewis 1974), it is safe to assume that any version of DAB which Lewis would have deemed worthy of serious consideration would be one where a single mental state can legitimately be interpreted in two ways: as a desire or as a belief.

Is DAB a putative conceptual truth or a normative requirement of some sort? We do not want to rule out violations of DAB as conceptual impossibilities. Indeed, Lewis (1989) is quite clear that the 'omniscient supervillain' is conceptually possible: someone who knows everything about what is good and still desires to do evil (see also his unpublished Lewis [1985] 2023). Likewise, it is perfectly possible that you understand that vegemite is a poor choice of ice cream flavour (Lewis 1988b) and yet desire it.

If we are merely asking whether it is *possible* for someone, at some time, to satisfy DAB then the answer is obviously that it is. But that is a hollow victory for

[4] One might worry that because credences are normalised and so have an upper bound, Desire as Belief entails that desires also have an upper bound. We take this to be unproblematic for our present purposes. First, this did not make Lewis reject DAB. Second, below we only defend DAB when restricted to second-order desires. Given this restriction, taking the first-order desire to have an upper bound allows for a notion of *desirability simpliciter* which Lewis himself relies on (we return to this in §3). Third, there are independent reasons to think that formal representations of desire should be normalised. See Sepielli (2012) for an argument.

defenders of DAB. The real question is whether agents *ought* to satisfy it. So, we take DAB to denote a putative requirement of practical rationality. While Lewis is not as clear on this point as one might have hoped, there is reason to think he would not object to that classification. It is a mistake, says Lewis, to think that moral beliefs automatically motivate the rational. But it is a substantive, not conceptual mistake. As he grants in (2005), it is not a violation of language to label DAB a requirement of rationality.[5] It is merely false: rationality does not require that your beliefs about the good match your desires—or so says Lewis. Now, on to his argument.

The central idea behind Lewis' objection is this: rational beliefs are sensitive to evidence in ways that desires are not. So, we can construct scenarios in which you initially satisfy DAB, but upon receiving evidence, you are rationally required to update your beliefs in a way that violates DAB. Why not just revise your desires as well, so as to continue satisfying DAB after the belief update? Because that would imply that how much you desire something should be a function of your evidence. And that is not generally plausible. Learning that it is raining outside has, by itself, no *normative* bearing on your desires. In particular, it does not constrain how much you *should* desire to get drenched.

This gives us the second step of Lewis' argument: if DAB holds, so should it for our conditional credences:

DACB: $\mathcal{V}(A) = \mathcal{C}(\mathring{A} \mid E)$.

Lewis focuses on cases where $E = A$, where indeed DACB seems plausible. Our beliefs about how good some state of affairs is should be independent of whether that state of affairs obtains. How good it would be to eat ice cream, sleep, or run an

[5] Lewis (1994) argues that rational requirements are constitutive of mental contents. It seems, therefore, that he would indeed classify Desire as Belief, if correct, as a rational requirement. Invoking such a broad sense of 'rationality' is not unfamiliar. It is similar to what Broome (2013) calls the 'central ought', which denotes the normative domain that encompasses putative requirements such as enkratic coherence or means–ends coherence. See Wallace (2020) for further context.

honest business is not influenced by whether one in fact eats ice cream, sleeps in, or is honest.

But then we can substitute identicals in DAB and DACB when $E = A$, and so we get the *independence* requirement:

IND: $\mathcal{C}(\mathring{A}) = \mathcal{C}(\mathring{A} \mid A)$.

Said differently, DAB and DACB jointly entail that beliefs about how good some proposition A is should be invariant under conditionalisation on the fact that the target state of affairs obtains. Learning that A is true should make no difference to your credence that A is good.

It is against IND that Lewis runs his objection. Suppose you learn the disjunction that $(A \vee \mathring{A})$. Learning this is some evidence that $\mathring{A}$ is true and so you rationally increase your credence in $\mathring{A}$. But $(A \vee \mathring{A})$ says nothing about how likely $\mathring{A}$ is in light of A, and so your conditional credence in $\mathring{A}$ given A should remain unchanged. So, after updating, your credences should satisfy the inequality $\mathcal{C}(\mathring{A}) > \mathcal{C}(\mathring{A} \mid A)$, thereby violating IND.

An example is in order.

> **Sophia:** As a devoted steward of principle, Sophia desires only what she believes to be good, and she always proportions her beliefs to the evidence. Sophia is considering whether to join the local theatre group. On the one hand, acting might be great fun. On the other, it might be anxiety inducing. Her trusted friend tries to motivate her and says: 'if you don't act, you'll miss out'. Being committed to classical logic, Sophia interprets her friend's testimony as material implication and so takes the testimony to say: either she acts or acting is good.

As should be clear, Sophia satisfies both DAB and DACB. She aligns her desires with her beliefs about the good and her beliefs reflect what her evidence tells her: by her lights, the goodness of $\mathring{A}$ is independent of whether the corresponding proposition A

obtains. Suppose Sophia trusts her friend and so becomes certain that either she acts or acting is good. Hopefully, it is both—she acts, and it is indeed good—but perhaps not. She can rule out the possibility that she never acts and misses out on a great good. How should Sophia update her beliefs? She should become more confident that acting is good. Perhaps she should not become much more confident than she already was. The friend's testimony was, after all, disjunctive and so in that sense unspecific. This need not concern us here. All we need is that her credence in the goodness of acting should increase at least a little.

However, her *conditional* credence that acting is good *given* that she acts should stay the same. This is so since the conditional credence is defined as a ratio of unconditional credences. In particular, $\mathcal{C}(\mathring{A} \mid A) = \mathcal{C}(\mathring{A} \wedge A)/\mathcal{C}(A)$. When Sophia learns that $(\neg A \supset \mathring{A})$, which is equivalent to $(A \vee \mathring{A})$, and in turn to $\neg(\neg A \wedge \neg \mathring{A})$, she rules out all the not-A-and-not-$\mathring{A}$ possibilities. But that does not influence the ratio of $(\mathring{A} \wedge A)$ possibilities to A possibilities. This is also intuitive: the friend said nothing about the goodness of acting given that she acts and so she has no reason to change her corresponding conditional credence. Thus, the rational response for Sophia is to violate the independence requirement, which is jointly entailed by DAB and DACB. DACB was motivated by DAB, and so they must both go. 'QED', says Lewis.

To summarise: we end in contradiction if we connect desires to beliefs and then insist that desires are not required to be responsive to evidence in the way beliefs are. However, the devil is in the details. As we will argue in §4, a certain set of desires are so constrained given a theory of value which links value with forms of epistemic idealisations. And Lewis' own theory of value is such a theory, as we will argue in the next section.

3 Lewis' dispositional theory of value

Lewis' metaethical theory of value consists of two main theses, respectively about valuing and value. For Lewis, *valuing* that A is to desire to desire that A. He does not offer a fully fleshed argument for classifying valuing as a desire-like state. However, the key to understanding why he classifies valuing as desire-like lies in the internalist aspect of his theory: under the right conditions, recognising the value of something comes accompanied with the dispositions to act accordingly (Lewis 1989: 113).

Lewis accepts the common division of mental states into belief-like and desire-like categories, where the latter is, as is typical, associated with intentional action and motivation. For Humeans generally, and so also for Lewis, this connection makes valuing better suited to the desire-like category (see Smith 1987). Of course, to identify valuing that A with a desire for A would be too strong. I may, for example, value a healthy diet and still not desire wheatgrass juice. Thus, to preserve the connection between valuing and action and motivation, Lewis proposes that valuing is desire to desire that A.[6] This formulation accommodates problematic cases: while I dislike wheatgrass juice, I do wish I were the kind of person who desires it, since I value a healthy diet.

It is worth pointing out that Lewis in his (1989) seems to treat desires as binary. That is, he seems to suggest that when I value A I have a second-order desire *simpliciter* to have a first-order desire *simpliciter* for A. But as he points out in his (1996), desires are graded attitudes, so the binary treatment must merely be a shorthand. This raises the question of how exactly Lewis' theory of value should be formalised. It is not enough to say that valuing A is a high second-order desire for a first-order desire, *to any degree*, for A. The first-order desire needs to be sufficiently

[6] Lewis (1989: 115) tentatively stipulates that value depends on second-order desires, rather than highest-order desires, to allow for the possibility of desiring different values. We follow this, though our proposed principle, Euthymia, could be straightforwardly amended to focus on highest-order desires.

high such that it can truly be said that one desires *simpliciter* that A.

In other words, if we define second-order desires as $\mathcal{V}(\mathcal{V}(A) = x) = y$, the question is what values x and y must take when someone values A. The numerical value of y will determine the degree they value A. But what about x? The formalisation is meant to capture a second-order desire *that one desires that* A, since this is how Lewis speaks of valuing. So, we tentatively propose defining a valuing attitude as a second-order desire for a maximal desire. That is, $\mathcal{V}(\mathcal{V}(A) = 1) = y$. However, it might turn out that valuing states need not have a maximal first-order desire as content, i.e. $x = 1$. If it turns out that $\mathcal{V}(\mathcal{V}(A) > .9)$, for example, suffices for desiring to desire *simpliciter* that A, then this is how valuing should be formulated. We leave this as an open question here.

So much for valuing. *Value*, Lewis says, is what we would desire to desire in ideal conditions. His view is subjectivist, meaning that what is valuable depends on us. However, he avoids reducing value to personal preferences, as we can value things that are not, in fact, valuable. In such cases, we desire to desire what we would not desire to desire in ideal conditions. For example, I might desire to desire immortality, but an ideal version of myself that considers all the possibilities of eternal life may not.

Unlike most ideal advisor theorists who require the advisor to be evidentially (Firth 1951) or rationally (Smith 1994) idealised, Lewis argues that the relevant kind of ideal self is ideal in only one respect: they are fully imaginative. For example, an imaginatively ideal agent never fails to consider the social repercussions of winning big at the lottery, and they never fail to draw a distinction (if such there be) between desire satisfaction and well-being when they consider the good life. It is exactly because an imaginatively ideal version of myself would *not* desire to desire to win, that winning is *not* valuable for me.

There is one final element of the view that needs unpacking. Lewis explains his internalism not just in terms of motivational states like valuing but also in terms

of value (1989). And it makes sense that he thinks the latter is relevant, too. Value is what an idealised agent would desire to desire. The desires of this ideal agent are motivationally relevant to us non-idealised people because *these agents are still us*, just in ideal conditions.

We are now in a position to see that the dispositional theory of value is vulnerable to Lewis' own argument against Desire as Belief. Lewis' view commits him to the claim that our beliefs about what is valuable necessarily go hand in hand with some of our second-order desires. Given his internalism about value, this causes trouble.

Consider the mental states we might have regarding value. A straightforward example is when we desire to desire what our ideal selves would desire to desire. However, there are other possible states we could have about value. When engaging in imaginative exercises, we form various attitudes about what our ideal selves would desire to desire, some of which are ordinary beliefs. For instance, I might come to believe that my idealised self would not desire to desire immortality.

This is a *belief* about what is valuable, if anything is. While such beliefs may be uncommon—arguably, we do not typically frame thoughts about what is valuable in these terms—they are nonetheless both possible and distinctively about value. Moreover, Lewis' internalism requires that these beliefs be connected to motivation, albeit not by a direct desire (Railton 2015: 540). Realising that my ideal self would desire to desire A may not move me, but under internalist assumptions, it seems plausible that it should. Indeed, there seems to be something wrong with someone who realises that a more ideal version of themselves would desire to desire A and yet remain wholly unmoved in their desire to desire that A. Lewis' ideal advisor theory makes this connection particularly salient, again, because they realise that *they themselves* would desire to desire A in idealised conditions.[7]

[7] The reasoning here parallels van Fraassen's (1984) Rational Reflection principle: you should defer to your future self since, all things being equal, you should expect your future self to be in a superior

Thus, even if our beliefs about our ideal selves do not always correspond to first-order desires, they ought to be accompanied by second-order desires. We should desire to desire what we believe our ideal selves would desire to desire. But what you believe your ideal self would desire to desire just is what you believe to be good. There is, then, a sort of necessary connection between our beliefs about what is good, understood as beliefs about our ideal selves, and our second-order desires.

The necessary connection can be understood in terms of practical rationality: we must, on pain of practical irrationality, desire to desire what we believe to be valuable. However, if beliefs about what is valuable necessarily go hand in hand with second-order desires, the dispositional theory runs into the same contradiction as the Desire as Belief did. The problem is that we have two different descriptions of our states that will necessarily go hand in hand to keep our thinking about what is valuable consistent.

One way to resist this argument is to question the role of these beliefs about our idealised selves in our thinking about what is valuable; perhaps all our thinking about value should be understood in terms of second-order desires only. One complication is that beliefs about our idealised selves are central to Lewis' view, since they allow for the possibility of truth and knowledge about value. Lewis says about his dispositional theory of value:

> Our theory makes a place for truth, and in principle for certain knowledge, and in practice for less-than-certain knowledge, about value. But also it makes a place for ignorance and error, for hesitant opinion and modesty, for trying to learn more and hoping to succeed. (1989: 123)

The key advantage of Lewis' theory over other forms of subjectivism is that it accommodates modesty. We can be wrong about what is truly valuable, and this allows for the possibility of being right in a substantial sense. Our views about what

epistemic situation.

is valuable can be true or false, and some can count as knowledge.

Lewis merely states that we can explain this given the idealised element of the view, but we can fill in the details. The question 'is A truly valuable?' can be substituted without loss by the question 'is A valuable?'[8] Lewis already interpreted the latter as one of whether we would still desire to desire that A in ideal conditions. Similarly, knowing that A is valuable, we might think, amounts to continuing to desire to desire that A even after passing a certain threshold of imaginative exercises. We return to this point in §4.1.

While this account of our valuing states may avoid positing two states that necessarily go hand in hand, since all we need to explain our thinking about what is valuable is our second-order desires, the account is not without problems. It commits us to counting some desires as beliefs, since truth and epistemic evaluability are properties we commonly associate exclusively with beliefs.

One could deny that a state being truth-apt and epistemically evaluable commits us to treating said state as a belief. Whether this response holds depends on the kind of metaphysics of belief one adopts. Under a relatively undemanding view like Lewis' interpretationism (1974), the claim would be hard to resist.

Indeed, interpretationist views—and more broadly on most functionalist views (which Lewis' view may be categorised as, see his 1994)—to be a belief just is to play certain roles. Truth and epistemic evaluability seem to be distinctive features of belief. While there are other truth-evaluable states, such as imaginings or suppositions, only beliefs seem subject to epistemic evaluation (for a similar point, see Beddor 2019). Thus, if valuing states are both truth-apt and epistemically evaluable, they effectively play the role of beliefs. This means that Lewis' own account of valuing requires us to understand at least some of our desires as beliefs.

However, if that is the case, following the argument against DAB, how should

[8] This substitution is more plausible given that Lewis endorses a form of deflationism about the truth predicate (see his 2001). However, the deflationist reading is not strictly needed.

the target state evolve in the face of evidence? When interpreted as a belief, it should update in accordance with conditionalisation. Not so when the state is interpreted as a desire. And so the dispositional theory of value leads to contradiction because the two interpretations cannot coherently describe the same state. As a result, if Lewis' argument against DAB is successful, the dispositional theory falls prey to the same issue.

4 A revised Desire as Belief thesis

The argument presented above constitutes good reason, we think, for Lewis to concede that Desire as Belief should not be rejected entirely, so in this section we put forth a revised version which constrains only second-order desires. Indeed, Lewis (1988a: 325) himself suggests that a plausible version of Desire as Belief could limit the range of desire-types the thesis quantifies over. Yet he never considers restricting to second-order desires given their belief-like properties and their connection to motivation.

The plausibility of constraining second-order desires is perhaps best appreciated when we recall that, for Lewis, having a second-order desire for something is *valuing* it. And this is something we can get right or wrong much like we can get beliefs right and wrong. One advantage of the following proposal is that it salvages the dispositional theory of value. The upshot is a theory of value that does constrain some desires rationally, but which leaves ample room for Lewis' scepticism about anti-Humeanism and the prospects of evaluating desires more generally.

When you believe that honesty is valuable, you judge that the ideal version of yourself would desire to desire to be honest. When next you lie, you do as you desire, not as you ideally desire to desire to do. Nothing inconsistent about it. But you cannot then rationally *believe* that your lying was good. After all, your beliefs about what is good just are your beliefs about what is valuable and so your beliefs about what you ideally desire to desire. This is the crux of the intuitive judgment

that Augustine cannot consistently claim that stealing pears is wrong while at the same time value stealing them.

Such a combination of attitudes reflects a form of incoherence. This mimics the kinds of incoherence involved in practical and epistemic akrasia, where your evaluative beliefs conflict with your intentions or first-order beliefs respectively. This suggests that it is a normative requirement that your beliefs about the good match what you value. More precisely, the degree to which you believe it would be good that A is the degree to which you should value that A obtains. Given the formulation of valuing as $\mathcal{V}(\mathcal{V}(A) = 1) = y$, we formulate this requirement as follows:

> **Euthymia:** For any agent with credence function $\mathcal{C}$ and desirability function $\mathcal{V}$, and for any proposition A and the corresponding evaluative proposition $\mathring{A}$, it is rationally required that $\mathcal{V}(\mathcal{V}(A) = 1) = \mathcal{C}(\mathring{A})$.

Euthymia says that a rational agent aligns their degree of belief that some proposition is good *simpliciter* with their degree of desire to desire *simpliciter* that the proposition obtains. Put more succinctly, Euthymia requires your beliefs about the good to match your values.

What if you are unclear about how desirable it would be to desire A to various degrees? For example, suppose your second-order desires for A divide between two possibilities: $\mathcal{V}(\mathcal{V}(A) = .3) = .6$ and $\mathcal{V}(\mathcal{V}(A) = .5) = .4$. Then it seems plausible that you should calculate the expected desirability of desiring A by taking a weighted average over your second-order desires, that is, .38.

We can generalise this reasoning in case your second-order desire divides between several possibilities: $\mathcal{C}(\mathring{A}) = \sum_x x\mathcal{V}(\mathcal{V}(A) = x)$.[9] The expected value represents how desirable it is, given your current evaluative perspective, to desire

[9] This parallels how we get the Principal Principle (Lewis 1980) to apply to rational agents even though they are rarely certain what the objective chances are. Taking the weighted average of the chance hypotheses with their credences as weights, you get the expected objective chance. Then the Principal Principle applies.

that *A simpliciter*. So, your evaluative credence should match this value. Given this, Euthymia has theoretical bite even if we do not typically form second-order desires for desiring *simpliciter* some proposition or form credences that these propositions are good *simpliciter*.

However, revising Desire as Belief does not make Lewis' objection go away. So, in the following we motivate the ways in which the revised thesis can escape it.

4.1 Staying enthymatic in the face of evidence

Can someone who satisfies Euthymia escape Lewis' objection? His counterexample involves gaining evidence which requires you to change your evaluative beliefs without changing your conditional beliefs about the target proposition. Euthymia is satisfied as long as your second-order desires correspond to your beliefs about the good. This correspondence can be maintained in the face of evidence in two ways: either neither your second-order desires nor your evaluative beliefs change at all, or else both your second-order desires and your evaluative beliefs change the same way.

Lewis' argument shows that the first strategy leads to contradiction: then evaluative beliefs cannot be probabilities. The second strategy was unmotivated when we focused on first-order desires. But, we will argue that there are good reasons to think evaluative beliefs and second-order desires should shift in tandem.

Let us start by taking a step back. In line with our earlier suggestion that second-order desires are all we need to explain our thinking about what is valuable, metaethical expressivists hold that, at a fundamental level, our thought about value is best understood as desire-like: its primary function is not to represent the world, as beliefs do, but to coordinate action. Yet expressivists also aim to accommodate the fact that moral judgments exhibit some surface-level belief-like features, such as their truth and epistemic evaluability. Some of the most developed expressivist accounts, moreover, link this kind of evaluability directly to the notion of stability in the face of incoming evidence.

We value many things, perhaps certain experiences, relationships, or achievements which help satisfy our desires. And we believe many things to be good. Call this an evaluative outlook: the set of mental states that capture an agent's evaluative beliefs and valuing attitudes. There are two features of evaluative outlooks that help explain how some desire-like states can be both stable and evaluable in terms of truth or knowledge.

Consider evaluability first. Many changes to our evaluative outlooks strike us as improvements. Lewis thought that the primary improvement to our evaluative outlooks consists in a greater imaginative capacity. Sophia's change from an evaluative outlook that values theatre acting moderately, to one that values it significantly, was an improvement insofar as she correctly judged the evidence presented to her.

To say that a series of imaginative exercises can serve as an improvement seems intuitive, even if the mental states in question are desire-like, such as second-order desires. Expressivists like Blackburn (1996, 1998) suggest that this idea of improvement can ground the kinds of evaluations we make of our second-order desires (see also Gamester 2022; Sinclair 2021: 193). These evaluations typically include knowledge and truth. To be sure, this core idea need not appeal exclusively to expressivists (see, for example, Railton 2017). Indeed, the idea is very close to Lewis' view: if our evaluative outlooks can improve, then there can be learning and knowledge, if our outlooks can worsen, there can be error and modesty.

Now consider stability. What role do concepts like truth or knowledge play in this picture? Whatever else we might say about them, a core feature of true beliefs—and beliefs that amount to knowledge—is that genuine improvement in our evidence, reasoning capacities, or epistemic standards leads to increased stability in the face of future evidence. This stability is what makes them apt to serve as standards in processes of belief revision. Expressivists argue that, given how we think about improvement, truth and knowledge can play a similar role when it comes to our evaluative outlooks.

The prospect of improvement is what makes it rational to revise our views in the face of evidence. Consider:

> **Felix:** The young transhumanist Felix longs for immortality. During his philosophy studies, he encounters Bernard Williams' *The Makropulos case*, which compellingly argues that immortality would turn out rather tedious. Disturbed, Felix revisits the argument again and again. Over time, he becomes less confident that immortality is so great.

The case, we take it, reflects a common and rational evolution of mental states in someone trying to improve their evaluative outlook. Suppose Felix initially has a rationally high credence that immortality is good and that he rationally lowers it in light of the evidence he receives. This much is straightforward: learning that immortality probably turns into misery indicates that one should lower one's credence in its goodness. Should Felix also change any of his desires?

It seems that he should. Evidence which improves his imagination brings him closer to his ideal self. Since Felix now becomes less confident that his ideal self desires to desire immortality, this puts pressure on him to lower his current second-order desire accordingly. In doing so, he can still satisfy Euthymia: his belief and second-order desire should shift in tandem. It seems plausible that this is the rational response in Felix' case. After all, it is the same piece of evidence (the Makropulos case) which challenged both his belief and desire, so changing the same way is the only non-arbitrary update available to him.

Can Felix escape Lewis' objection? Recall that the original argument assumes DAB and then argues that the conditional version of DAB, namely DACB, also holds. These jointly entail the problematic independence requirement. However, once we confine our focus to second-order desires, DACB should be rejected. We can formulate it like this:

> **Conditional Euthymia:** For any agent with credence function $\mathcal{C}$ and desirab-

ility function $\mathcal{V}$, and for any proposition A and the corresponding evaluative proposition $\mathring{A}$, it is rationally required that $\mathcal{V}(\mathcal{V}(A) = 1) = \mathcal{C}(\mathring{A} \mid E)$.

When $E = A$, this implies that A is evidentially irrelevant to how much one desires to desire A. This opens the door for Lewis' objection.

However, we should question the plausibility of Conditional Euthymia. The principle presumes imaginative completeness which non-ideal agents lack. Such agents, if rational, are modest: they do not assume that their current evaluative stance is stable across all possibilities since they recognise that their current second-order desires may misalign with what is valuable. Thus, such agents should acknowledge that they may be mistaken about what is good and that some evidence, including evidence of the form $(A \vee \mathring{A})$, is relevant to refining their evaluative outlooks. Accordingly, when the credence in $\mathring{A}$ increases in response to such evidence, a rationally modest agent should also increase their second-order desire, preserving the alignment demanded by Euthymia. This is because the evidence which informs whether $\mathring{A}$ also informs the agent what their ideal selves desires to desire.

Apply this to Felix: given his imaginative shortcomings, Felix should not take himself to be able to evaluate immortality conditional on each of the many distinct ways in which immortality could be realised. In this way, Felix should define $\mathcal{C}(\mathring{A} \mid E)$ for various E which are not equal to his current second-order desire, including for $E = A$, thereby contradicting Conditional Euthymia. This is not because he discovers that immortality is bad given that it occurs, but because he rationally doubts his own ability to get his evaluative judgments exactly right.

In appealing to the non-ideality of Felix, does the solution thereby only work for imaginatively non-ideal agents? Arguably not. Suppose we stipulate that Sophia is imaginatively ideal:

> **Ideal Sophia:** Sophia's counterpart is also uncertain about the value of acting. And she, too, rationally apportions her beliefs in accordance with the available

> evidence. The relevant difference is that ideal Sophia is imaginatively ideal: through vivid imagination, her epistemic space has become as rich as can be: she is fully acquainted with the realisation of every possibility there is.

For any possible state of affairs, ideal Sophia is well acquainted with what that would be like. As such, she has already determined how much she would desire to desire any one of those states, and she has also formed evaluative beliefs about their goodness. It might therefore seem natural to conclude that her evaluative outlook, having been shaped through exhaustive imaginative engagement, should be fully stable under further evidential updates.

However, even if Sophia's imaginative capacities are ideal, we need not assume that she regards her evaluative judgments as infallible. For example, the revelation from an oracle should plausibly rationalise a change in her beliefs about the good. This will be so, even if the oracle's revelation is in fact misleading. Although this revelation will not serve to alter which possibilities Sophia considers, it should arguably put pressure on what she values.

After all, just because Sophia is imaginatively ideal, she might harbour some doubt as to whether she has located the correct evaluative profile. If so, she should not assume that her current degree of confidence that acting is non-instrumentally good is maximally resilient across all possible evidential updates, to use Skyrms's (1980) terminology. To do so would be to treat her evaluative credence as beyond evidential revision. But that is exactly the kind of immodesty which Lewis' theory of value was meant to rule out.

Accordingly, ideal Sophia cannot rationally dismiss *all* potential evidence. She should allow some evidence to bear on her confidence that acting is good. This does not reflect a failure of imagination, but rather a form of rational humility about the correctness of her evaluative judgments. Evidence whose content bears directly on whether $\mathring{A}$ is true—revelation from an oracle, for example—must be evidentially relevant for Sophia. In light of this, Conditional Euthymia loses its plausibility even

for imaginatively ideal agents. Ideal Sophia should not take her evaluative outlook to be immune to revision conditional on every proposition, including $\mathring{A}$ itself. Rejecting this assumption blocks Lewis' argument since the objection relies precisely on such conditional stability.

The point is not that every piece of evidence must be treated as relevant, but that evaluative beliefs must remain responsive to some evidence if they are to allow for rational improvement at all. For example, Sophia may or may not be open to revising her evaluative belief in the face of testimony from putative experts, depending on whether she takes such testimony to introduce genuinely new evaluative considerations. In virtue of being imaginatively ideal, Sophia might maintain that whatever she learns from an esteemed philosopher or her trusted friend, it is not that $\mathring{A}$. Perhaps she regards any such testimony as merely reporting a different evaluative stance. Williams' claim that immortality would be a bore does not introduce a possibility to Sophia that she had failed to consider, as it did for Felix.[10] Still, if she is rational, she will remain open to revising her evaluative credence $\mathcal{C}(\mathring{A})$ if she were to learn that $(A \vee \mathring{A})$, for example. Such openness just is to reject Desire as Conditional Belief.

Desire as Conditional Belief may be plausible when applied to first-order desires like those about the desirability of getting drenched by rain. But it fails for second-order desires in rationally modest agents. So long as we reject the principle at the second-order level, Euthymia survives Lewis' objection.

[10] Such disagreement among putatively ideal agents is evidence of non-convergence. Lewis (1989) makes room for this possibility. If there is non-convergence, then evaluative truths should be indexed to (groups of like-minded) individuals. So, when others report valuing theatre or immortality more highly, this may be perfectly true for them without generating any rational pressure on Sophia to revise her stance. See Egan (2012).

5 Two alternative solutions

We end by contrasting our proposal with two prominent solutions offered in the literature which promise to salvage the original DAB thesis. Our aim here is not to offer a decisive refutation of these views, but to highlight the costs they incur relative to Euthymia.

First, Hájek and Pettit (2004) respond to Lewis' objection by exploiting what they call the 'indexicality loophole'. On their view, the proposition $\mathring{A}$ does not retain a fixed content across time. Instead, its meaning shifts with the agent's evaluative perspective. This fits well with well-known versions of subjectivism according to which value depends on your current mental states: when your mental states change, what is valuable changes accordingly.

For example, Felix initially holds that immortality is good and desires it accordingly. After reading Williams, he comes to doubt the value of immortality. His credence in $\mathring{A}$ decreases, while his standing desire for immortality remains unchanged, so $\mathcal{C}(\mathring{A}) < \mathcal{V}(A)$. Hájek and Pettit propose that Felix now instead adopts a credence in a new proposition, $\mathring{A}^*$, such that $\mathcal{C}(\mathring{A}^*) = \mathcal{V}(A)$. As a result, even when an agent rationally lowers their credence in $\mathring{A}^*$, they can still satisfy Desire as Belief thesis so long as $\mathring{A}^*$ now expresses a concept 'good*' which has a different meaning than 'good'.

The structural form of Desire as Belief is thereby preserved. However, this is only because the relevant evaluative proposition has changed. The solution amounts to reinterpreting the contents of beliefs to preserve isomorphism with desire, rather than offering a substantive explanation of why belief and desire should align. The upshot is an extreme conceptual pluralism: every time one updates an evaluative belief, one also adopts a new evaluative concept. Felix does not learn that he was mistaken about what is good, he simply shifts to a different standard. It seems that this undermines evaluative improvement. Felix is no longer correcting a mistaken

evaluative belief but merely talking past his former self.

This is worrying enough as it is, but it also leads to a further problem. If each evaluative update involves adopting a new concept of 'good', then agents might become vulnerable to a kind of diachronic irrationality. Suppose Felix, before reading Williams, is willing to pay £100 to secure immortality, given his high credence that it is good. After updating, he values immortality less and so would pay less. But if his value concept has changed, he cannot even say that he regrets the earlier choice; he simply has a different notion of value.

A clever bookie could exploit this: selling Felix the promise of immortality while his $\mathcal{C}(\mathring{A})$ is high and buying it back after his $\mathcal{C}(\mathring{A}^*)$ drops. Felix is guaranteed to lose money due to his shifting evaluative concepts. The possibility of such a money pump suggests that the indexicality loophole licences patterns of decision-making that are predictably exploitable and thereby practically irrational (for an early argument that predictable exploitability is a mark of irrationality, see Lewis (1999)).

Second, Bradley and List (2009) offer a different solution by sharply distinguishing between evaluative and descriptive propositions. To preserve Desire as Belief, they posit two separate probability functions—one for evaluative propositions and one for descriptive ones—which operate independently. Purely evaluative propositions will be invariant under conditionalisation on non-evaluative propositions and vice versa. This avoids Lewis' problem case since neither of the two probability functions are defined over 'mixed evidence' like $(A \vee \mathring{A})$ which mixes the evaluative and the descriptive.

While the solution works, it seems to undermine the need for a solution in the first place. The motivation for Desire as Belief is to capture a rational connection between what we believe to be good and what we desire. That connection becomes opaque if beliefs and values are assigned and evolve independently. If value and fact are systematically unrelated, what sense remains in claiming that some desires are belief-like? Indeed, their radical subjectivist solution is quintessential Humean: there

is *no* interaction between desire and belief. As such, it is normatively inert possibility that an agent happens to satisfy Desire as Belief. Since it undermines the possibility of evaluative improvements, it is difficult to imagine that someone like Lewis would be tempted by this solution.

These contrasts help clarify what is at stake in our proposal. Euthymia preserves a meaningful normative link without resorting to indexical shifts or dual belief systems. It holds fixed the content of 'good' and models how rational agents adjust both evaluative belief and second-order desire in light of evidence. Felix should revise his desire to desire immortality as he revises his belief about its value, preserving at least some of the spirit, not just the letter, of Desire as Belief.

6 Conclusion

Few of us are lucky enough to desire only what we believe to be good. While this may reflect a psychological failure, Lewis denies that it amounts to a rational one, and so he objects to Desire as Belief. Nevertheless, as we have argued, Lewis cannot reject it wholesale due to his commitments to cognitivism about evaluative judgments and internalism about value.

However, we have also shown that it *is* possible to link second-order desires and evaluative beliefs in a way which avoids the objection. Indeed, we think Lewis, along with those who link value with forms of epistemic idealisations, should welcome such news. Euthymia strikes the balance between Humean scepticism about general norms for desires while preserving the possibility of evaluative improvement through evidence responsiveness.

References

Beddor, Bob (2019) 'Noncognitivism and Epistemic Evaluations'. *Philosophers' Imprint*, 19.10, pp. 1–27

Bex-Priestley, Graham (2018) 'Error and the Limits of Quasi-Realism'. *Ethical Theory and Moral Practice*, 21.5, pp. 1051–1063

Blackburn, Simon (1996) 'Securing the Nots: Moral Epistemology for the Quasi-Realist'. In: *Moral knowledge? New readings in moral epistemology.* Ed. by Walter Sinnott-Armstrong and Mark Timmons. OUP, pp. 82–100

— (1998) *Ruling Passions: A Theory of Practical Reasoning.* New York: OUP

Bradley, Richard and Christian List (2009) 'Desire-as-Belief Revisited'. *Analysis*, 69.1, pp. 31–37

Broome, John (2013) *Rationality Through Reasoning.* Wiley-Blackwell

Egan, Andy (2012) 'Relativist Dispositional Theories of Value'. *Southern Journal of Philosophy*, 50.4, pp. 557–582

Firth, Roderick (1951) 'Ethical Absolutism and the Ideal Observer'. *Philosophy and Phenomenological Research*, 12.3, pp. 317–345

van Fraassen, Bas (1984) 'Belief and the Will'. *Journal of Philosophy*, 81.5, pp. 235–256

Gamester, Will (2022) 'Fallibility Without Facts'. *Ergo*, 8.40

Gregory, Alex (2021) *Desire as Belief: A Study of Desire, Motivation, and Rationality.* OUP

Hájek, Alan and Philip Pettit (2004) 'Desire Beyond Belief'. *Australasian Journal of Philosophy*, 82.1, pp. 77–92

Lewis, David (1974) 'Radical Interpretation'. *Synthese*, 27, pp. 331–344

— (1980) 'A Subjectivist's Guide to Objective Chance'. In: *Studies in Inductive Logic and Probability, Volume II.* Ed. by Richard Jeffrey. University of California Press, pp. 263–293

— (1986) *On the Plurality of Worlds.* Oxford: Blackwell Publishers

— (1988a) 'Desire as Belief'. *Mind*, 97.418, pp. 323–32

— (1988b) 'What Experience Teaches'. In: *Proceedings of the Russellian Society.* University of Sydney, pp. 29–57

Lewis, David (1989) 'Dispositional Theories of Value'. *Aristotelian Society Supplementary Volume*, 63.1, pp. 89–174

— (1993) 'Evil for Freedom's Sake'. *Philosophical Papers*, 22.3, pp. 149–172

— (1994) 'Reduction of Mind'. In: *A Companion to the Philosophy of Mind*. Ed. by Samuel D. Guttenplan. Blackwell, pp. 412–431

— (1996) 'Desire as Belief II'. *Mind*, 105.418, pp. 303–13

— (1999) 'Why Conditionalize'. In: *Papers in Metaphysics and Epistemology*. CUP, pp. 403–407

— (2001) 'Forget About the 'Correspondence Theory of Truth''. *Analysis*, 61.4, pp. 275–280

— (2005) 'Quasi-Realism is Fictionalism'. In: *Fictionalism in Metaphysics*. Ed. by Mark Eli Kalderon. OUP, pp. 314–321

— ([1985] 2023) 'Mass and Value'. In: *Philosophical Manuscripts*. Ed. by Frederique Janssen-Lauret and Fraser Macbride. OUP, pp. 181–183

Railton, Peter (2015) 'Lewis on Value and Valuing'. In: *A companion to David Lewis*. Ed. by Barry Loewer and Jonathan Schaffer. Wiley-Blackwell, pp. 533–548

— (2017) 'Learning as an Inherent Dynamic of Belief and Desire'. In: *The Nature of Desire*. Ed. by Federico Lauria & Julien Deonna. OUP, pp. 249–275

Schroeder, Tim (2020) 'Desire'. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer. Metaphysics Research Lab, Stanford University

Sepielli, Andrew (2012) 'Normative Uncertainty for Non-Cognitivists'. *Philosophical Studies*, 160.2, pp. 191–207

Sinclair, Neil (2021) *Practical Expressivism*. OUP

Sinhababu, Neil (2017) *Humean Nature: How Desire Explains Action, Thought, and Feeling*. OUP

Skyrms, Brian (1980) *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. Yale University Press

Smith, Michael (1987) 'The Humean Theory of Motivation'. *Mind*, 96.381, pp. 36–61

Smith, Michael (1994) *The Moral Problem*. Cambridge, Mass., USA: Blackwell

Wallace, R. Jay (2020) 'Requirements of Reason'. In: *The Routledge Handbook of Practical Reason*. Ed. by Ruth Chang and Kurt Sylvan. Routledge