

HUME, THE *BAD PARADOX* AND VALUE REALISM

One plausible interpretation of Hume’s famous claim that “reason is, and *ought* to be, the slave of the passions” holds that desires are necessary to motivation, that beliefs alone cannot motivate. Now some have held, against Hume, that there are in fact beliefs which could motivate all by themselves—namely, beliefs about what is good. For there to be such beliefs then cognitivism about value—the thesis that there are truth-bearing propositions about value—would have to be a live option. A recent slew of arguments, if sound, would demonstrate that such a cognitivism about value is in conflict with fundamental principles of practical rationality. While these proofs started life as an attack on a particular brand of anti-Humeanism in fact they have much broader application. Indeed, they threaten absolutely any cognitivist theory of value. They seem to force the cognitivist to deny core tenets of our best theory of rational decision making—decision theory—itself an articulation of the broadly Humean thesis that a rational action is one which serves one’s desires according to one’s beliefs. That would indeed be bad news for those who wish to locate their value realism within a broadly naturalistic understanding of the universe. Fortunately for value realists, the bad news turns out to be good news after all.

1 DISSONANCE

It is a familiar and depressing fact that our desires do not always conform to our beliefs about the good. Cognitivism about value seems well-placed to accommodate this familiar phenomenon, which might aptly be called *dissonance*. According to the cognitivist there are truth-bearing propositions about the value of things, and hence there can be genuine beliefs taking these value-propositions as their objects. Desiring may also be a propositional attitude, but it seems quite different from believing, and this is just what we need to provide the required conceptual space for dissonance. If believing and desiring are different kinds of mental attitude, it should come as no surprise that one’s desires might fail to be appropriately aligned with one’s beliefs about what is valuable. Various forms of non-cognitivism, on the other hand, hold that what purport to be expressions of beliefs about value are really just expressions of desire or approval. If that is right it is difficult to see how dissonance is really possible.

Realism about value is stronger than cognitivism. One might be a cognitivist about value, acknowledging that there are genuine propositions about value, but reject the idea that any of these propositions are really *true*. This familiar enough position is *nihilism*. (The most promising line for the nihilist, in my view, is to maintain that the class of interesting or substantive propositions about value are all truth-valueless rather than false.) The nihilist

could make good use (or anyway, *some* use) of the cognitivist account of valuing, but she will find it harder than the realist does to motivate *harmony*: having one's desires aligned with one's beliefs about the good. For the realist it is easy to provide both conceptual space, and motivation, for harmony. The realist takes there to be genuine truths about what is valuable. Consequently the realist thinks there are two regulative ideals which should constrain our valuing. First, one's beliefs about the good should be true. Second, one should desire things in proportion to their value. Elsewhere I have called these the regulative ideals of *truth* and *purity*. And it is pretty obvious that these two ideals together enjoin harmony.

2 THE *BAD* PARADOX

While the metaphysics of value realism is far from uncontroversial, the logic of valuing seems relatively straightforward. Despite these appearances, a series of recent arguments (beginning with David Lewis's 1988) would expose cognitivism (and hence realism) as logically problematic. They effectively show that the regulative ideal of harmony is in deep conflict with canons of practical rationality.

Those who have been following the DAB literature might find my characterization here more than idiosyncratic. The proofs started life as an attack on a particular *brand* of anti-Humeanism: that some beliefs necessarily motivate. This anti-Humean thesis is sometimes labelled the *desire-as-belief* thesis, or DAB. However, it could more accurately be labelled the *belief-as-desire* thesis, or *BAD*, since the claim is that a certain sort of belief (a belief about goodness) plays essentially the same role in motivation as a desire. Whatever acronym we settle on, these proofs have sought to show that there is something fundamentally problematic about *BAD*, because it conflicts with tenets of rational decision theory. Call this the *BAD* paradox.

What has not been widely appreciated, and what sparked my interest in this area, is that if these proofs are sound then they are bad news not just for the belief-as-desire thesis (which embodies a very strong version of internalism), but for cognitivism about value in general, and hence for any kind of realism about value. This is because they attack *BAD* by attacking the thesis that there could be an appropriate, systematic correlation—not just identity or a necessary connection—between beliefs about goodness and desire. If the proofs are sound, and the explications on which they are based are accurate, then the cognitivist will have to admit that it is virtually impossible for a rational agent to maintain harmony between his beliefs about the good and his desires. (We knew it was hard, but we never suspected that it would conflict with basic canons of rationality.) For the realist it is especially bad news, because it follows that a rational agent will almost certainly be deficient either on the score of

truth or on the score of purity. That's very bad.

I myself am not particularly interested in defending the very strong internalist thesis that there are a bunch of beliefs which *necessarily* motivate action. That is much stronger, for example, than the thesis that there be *some* internal connection between beliefs about value and desire.¹ I am, however, interested in showing that the realist can embrace what appears to be a very weak consequence of that thesis, the bare logical possibility of harmony, without rejecting rational decision theory.

In an earlier paper on this paradox I argued that conditionalization is a problematic method for updating credences when applied to changing propositions—propositions the truth values of which change over time. If the value of a state is a changing magnitude, then the earliest version of the *BAD* paradox seems to require a problematic application of conditionalization to changing propositions.² While I still think that the problems associated with change and conditionalization have not yet been adequately appreciated or solved, it turns out that simpler versions of the *BAD* paradox can sidestep the problem of changing values. They also make it much clearer what is at stake.

3 SIMPLE HARMONY

There are two rather different versions of the *BAD* paradox: a simple version and a more realistic version. In this section I show why I think the simple version involves an inadequate analysis of harmony (or belief-as-desire).

In what follows I will assume that belief and desire both take propositions as objects. If an agent wants to keep all his promises, and believes he will do so, then the object of his belief and his desire is one and the same—the proposition that he will keep all his promises. I will depart from most presentations of belief-as-desire by distinguishing between the properties of *goodness* and of *desiredness*. I assume that these are distinct features which the DAB principles attempt to bring together into a relationship of some interest to the cognitivist.

For the purposes of the analysis we begin with the familiar idea that a proposition

¹See, for example, Lewis 1989, for an account which does establish connections but of a very *iffy* sort.

²Oddie 1993, Lewis 1988.

determines a class of possible worlds—the *range* of the proposition. We need not accept the coarse-grained view that logically equivalent propositions, those with the same range, are identical. But we will assume for the analysis that a rational constraint on desiring (believing) is that logically equivalent propositions are desired (believed) to the same degree.

According to the cognitivist there are genuine propositions about what is and is not good. So one simple way of spelling cognitivism out might be this:

Simple Cognitivism

For each proposition A there is the proposition \hat{A} : that A is good.

Since harmony amounts to the alignment of one's desire that A with one's belief in the goodness of A (viz belief in \hat{A}) the following might suggest itself: one exhibits harmony (with respect to A) just in case the strength of one's desire for A is equal to the strength of one's belief in the goodness of A . That is to say:

Simple Harmony (or Simple *Belief as Desire*)

The degree of one's desire for A is the degree of one's belief that A is good:

$$D(A) = C(\hat{A}).$$

This is indeed a simple explication and one might wonder whether this really is the correct explication of the thesis, informally characterized as *desiring A just to the degree that one believes A good*. The italicized phrase is ambiguous. Does *degree* attach to the *belief* or to the *goodness* of what is believed? If we take the former reading we get: *desiring A to the same degree as the degree of belief in the proposition that A is good*. If we take the latter then we get: *desiring A to the same degree as the degree of goodness one believes (or estimates) A to have*. The explication given settles firmly on the former reading. But the latter reading seems just as natural, if not more so. What if one is pretty near certain that A is a little better than so-so. Then one invests 0 probability in the proposition that A is (thoroughly) good. But one's desire for A may not be 0. We will return to this problem.

To illustrate the paradox which this interpretation of belief-as-desire engenders, here is a short story.³ Frederic, who has been in a bonded contract since birth to serve the pirates, has made two promises: one to the pirates (that he will honor the contract according to which he will remain single and bonded in service to the pirates until his 21st birthday); and

³With apologies both to Gilbert and Sullivan!

a second to the love of his life (that he will marry her as soon as he is released from his contract). After making these promises Frederic learns firstly, that he was born on the 29th of February, making his 21st birthday coincide with the start of his 85th year of life; and secondly that the pirates are about to make a decision in a secret meeting about whether to release him from his contract early; and thirdly, that he must shortly settle on a wedding date in ignorance of the pirates' decision. Frederic is a Kantian, a slave of duty. In this situation his sole goal is *keep all his promises*. This is the only outcome he regards as good.

So for Frederic there are four salient outcomes:

Pirates release early & Frederic marries early (promises kept - GOOD).

Pirates release early & Frederic marries late (promise broken - BAD).

Pirates don't release & Frederic marries early (promise broken - BAD).

Pirates don't release & Frederic marries late (promises kept - GOOD).

According to Frederic, then, an early marriage (A) is a good thing (\hat{A}) if and only if the pirates decide to release him early.

Let's make a couple of assumptions from probability and decision theory. First, that it is rational to update one's beliefs by conditionalization on incoming information. Second, that degree of desire for any proposition is the expected desiredness of the outcome conditional upon that proposition. The object of the belief-as-desire proofs is to show that anyone who obeys simple harmony will end up flouting one or other of these two principles, both of which are rationally permissible even if not rationally obligatory.

Suppose all four outcomes have a positive initial probability for Frederic. For example he may think it 50/50 that the pirates will release him early, and is as yet undecided how to act. So in particular, both $C(\hat{A} \sim A)$ and $C(\sim \hat{A} A)$ are non-zero. Suppose that Frederic obeys simple harmony in the face of any old incoming information. Before making his decision whether to opt for an early or a late wedding, Frederic learns (from some apparently authoritative source) that if he marries early that will not be a good thing: $[A \rightarrow \sim \hat{A}]$. He updates C to C^+ by conditionalization on this information, and updates D to D^+ in conformity with the principle of desiredness. We then have (by simple harmony):

$$D^+(A) = C^+(\hat{A}) = C(\hat{A} | A \rightarrow \sim \hat{A}) = C(\hat{A} \sim A) / C(A \rightarrow \sim \hat{A}).$$

Since $C(\hat{A} \sim A) > 0$, and $\hat{A} \sim A$ entails $(A \rightarrow \sim \hat{A})$, $C(A \rightarrow \sim \hat{A}) > 0$, and so

$$D^+(A) > 0.$$

Where $C_p(Q)$ is $C(Q|P)$ and D_p is the desiredness function based on C_p , the principle of desiredness tells us quite generally that $D(P) = D_p(P)$, and so, $D^+(A) = D^+_A(A)$. By simple harmony, $D^+_A(A) = C^+_A(\hat{A})$. Hence:

$$D^+(A) = C^+_A(\hat{A}) = C(\hat{A}|A(A \rightarrow \sim \hat{A})) = C(\hat{A} \sim \hat{A})/C(\sim \hat{A}).$$

Since $C(\sim \hat{A}) > 0$ and $C(\hat{A} \sim \hat{A}) = 0$,

$$D^+(A) = 0.$$

Thus our principles jointly entail a contradiction. Frederic cannot obey both conditionalization and desiredness and keep his desires in harmony with this beliefs about goodness.

This argument is, of course, quite general and has nothing to do with Frederic's peculiar situation. In addition to the three principles the only substantive assumption is that $C(\hat{A} \sim A) > 0$ and $C(\sim \hat{A} A) > 0$.⁴

Having granted the existence of the propositions about goodness, the cognitivist should then be committed to the bare possibility of harmony. But that leads immediately to contradictions with decision theory, given modest assumptions. At least to a committed decision theorist, this looks like a very powerful argument for Wittgenstein's claim that "propositions of ethics are impossible". Well, even if not impossible, then at least somewhat paradoxical.

4 A PARALLEL ARGUMENT

If this argument is sound then it might seem to impugn more than the cognitivist's cherished property of goodness. Consider the non-cognitivist's cherished property of *desiredness*. Shouldn't it be possible for a perfectly rational agent to maintain harmony between his desires and his beliefs *about his desires*? Or to put this slightly differently, couldn't someone desire things just to the degree that he believes them to be *desired by him*? Call this *desire harmony*. That should be *possible*. Difficult, for sure, but *possible*. It would seem to be a regulative ideal for the non-cognitivist that she become thoroughly familiar with her own desires (*Know Thyself!*). So it had better not be a consequence of decision theory that it is impossible for an agent to know what her own desires are. (Indeed, it is essential to certain versions of decision theory that desires all be excogitable from one's dispositions to bet.)

⁴For rather similar simplified proofs, see Lewis 1996, and Byrne and Hajek 1997.

Now substitute *desired* throughout for *good*. Apart from the cognitivist postulation of a property (labelled “goodness”), the only principle required to generate the problem is the simple version of harmony. If simple harmony were the correct explication of the informal characterization it seems to follow by a parallel argument that conditionalization, expectation and desire-harmony also generate a paradox.

Let’s go through the parallel argument more carefully. We start with each *total* outcome u in Frederic’s small possibility space being either desired (desiredness=1) or undesired (desiredness=0). We extend desiredness to propositions in general by the usual principle. Now, in analogy with simple harmony, we make the assumption that there is property of *being desired* which each proposition either has or lacks: A^∂ . We impose the constraint of desire harmony, that our agent desires A just to the extent that he believes that A is desired:

Desire Harmony

The degree of one’s desire for A is the degree of one’s belief that A is desired:

$$D(A) = C(A^\partial).$$

The parallel argument concerning desire-harmony would require both that $C(A^\partial \sim A) > 0$ and $C(\sim A^\partial A) > 0$. That means our agent at the outset would have to be uncertain about whether or not A is desired. Thus he would have to be, to some extent, ignorant of his own desires. So the parallel proof would only go through on the assumption that our agent does not quite know himself after all. Now that hardly impugns the regulative ideal *Know Thyself!*. Rather, it reinforces it. Any agent who does not know his own desires perfectly will get into conceptual trouble. Nevertheless, our earlier misgivings about the explication of harmony leap out with greater force. Maintaining what I have called desire-harmony is simply having correct beliefs about one’s own desires. That is, believing that A has the very degree of desiredness that A actually has. But this is equivalent to: *desiring A to the same degree as the degree of desiredness one believes A to have*. That is, it is the analogue of the second, rejected, interpretation of the informal version of harmony. So, it seems we have erred by substituting a coarse-grained property of *being desired* for fine-grained degrees of desiredness, and then attempting to make up the deficit through fine-grained degrees of belief. Why not stick with degrees of desiredness for each proposition, and characterize harmony in terms of propositions which attribute those degrees?

5 HARMONY REVISED

We have isolated an awkwardness in the formulation of belief-as-desire. There is something wrong with harmony so characterized, whether it involves goodness or desiredness, and it appears to have something to do with the obliteration of degrees of the feature at issue, and the concomitant substitution of degrees of belief.

Goodness, like desiredness can come in degrees—not only the goodness of total outcomes, but of states and propositions generally. Even if the total outcomes are all either good or bad *simpliciter*, other states might be more or less good or bad, depending on the tightness of their connections with the good and the bad outcomes. Let $V(A)$ be the degree of goodness, or value, of A . (' V ' is often used in the literature to denote *desiredness*. It is important for the bearing of the argument on realism, however, that V and D be kept apart.) Once we have replaced the simple monadic attribute of goodness with the magnitude *value*, we can replace our first simplistic account of cognitivism.

Cognitivism

For each A and i there is a proposition $V(A)=i$ (or \dot{A}_i , for short) that A 's value is i .

Of course, cognitivism so characterized could still be appropriated by anyone who thought propositions or states of affairs admit of degrees of goodness. It does not commit us to any particular account of the value of states of affairs. Goodness could be reducible to something ontologically more basic, it could be mind-dependent, it could be projected. Even the nihilist might agree that there *are* such propositions.

For each degree of value an agent (an ideally rational one, of course) invests some credence in the proposition that A is valuable to that degree. It may be that a supremely opinionated agent will single out just one of the the degrees of goodness and invest maximal credence in the proposition that A is valuable to that degree. In that case there is a single proposition which fully represents the agent's beliefs about the value of A , and in that case harmony would consist in the agent desiring A exactly to the degree of value that she assigns with complete confidence to A . Call this:

Stringent Harmony

The degree of one's desire for A is the value one is certain A possesses:

$$D(A)=i \text{ iff } C(\dot{A}_i)=1.$$

While this doesn't have the right form to generate the paradox, it is clearly inadequate as a

general explication of the state of harmony. It may well be possible for an agent not to know exactly what his *desires* are, but on a realist account of value one is *typically* ignorant of exactly where the *value* of any particular state lies. So for all but the most radically opinionated, there won't usually be any *single* possible value for A in which a rational agent invests maximal credence. In that case, what is her *epistemic attitude* to the value of A, that attitude which should be harmonized with her desires?

Her epistemic attitude is most fully captured, of course, by the *spread* of her credences over the various possible values of A. But that is too complex a state for the strength of her desire for A to match. If harmony is to be characterizable the agent's epistemic attitude to A will have to be reduced to a single number, and harmony will presumably obtain if she desires A just to that degree.

Whenever one distributes credences over the value of a magnitude X, one reduction of that state to single number, is the *credence-mean estimate* of the magnitude. Let us just call this the agent's *epistemic estimate* (Est) of X.

$$\text{Est}(X) = \sum_i C(X=i)i.$$

We can represent an agent's epistemic attitude to the value of A by credence-mean estimate .
So:

$$\text{Est}(V(A)) = \sum_i C(V(A)=i)i = \sum_i C(\hat{A}_i)i.$$

We now characterize harmony as desire being identical to epistemic estimate of value.

Harmony

The degree of one's desire for A is one's estimate of A's value:

$$D(A) = \text{Est}(V(A)) = \sum_i C(\hat{A}_i)i.$$

This formula has been guided by the informal idea that one's desire should be in step with one's epistemic attitude to value, together with the independently motivated idea that, quite generally, epistemic attitude to a magnitude can be appropriately captured by C-mean estimate.⁵

⁵Lewis in his 1988 gives this same characterization in the general case where the value of propositions admits of degrees, although he does not plod through the reasoning I have just given. If he subscribes to it then he must have regarded it as too obvious to mention.

6 PARADOX REGAINED

By harmony we have:

$$D(A) = \sum_i C(\dot{A}_i)i$$

and $D^+(A) = \sum_i C^+(\dot{A}_i)i$

Assume there is an A such that

$$(\#) \quad C(A\dot{A}_0) > 0 \text{ and for at least one } i > 0, C(\sim A\dot{A}_i) > 0.$$

Consider an analogue of the proposition we used in the last simple proof: $E = [A\dot{A}_0]$, that if chooses A, that choice is of minimal value. Note that since $[\sim A\dot{A}_i]$ entails E, and $C(\sim A\dot{A}_i) > 0, C(E) > 0$. . Suppose our harmonious agent learns just $[A\dot{A}_0]$ and updates to C^+ appropriately.

$$\begin{aligned} \text{Then} \quad D^+(A) &= \sum_i C^+(\dot{A}_i)i && \text{by harmony} \\ &= \sum_i \{C(\dot{A}_i E)/C(E)\}i \\ &= \sum_{i=0} \{C(\dot{A}_i E)/C(E)\}i + \sum_{i>0} \{C(\dot{A}_i E)/C(E)\}i \\ &= 0 + \sum_{i>0} \{C(\dot{A}_i [A \rightarrow \dot{A}_0])/C(E)\}i \\ &= \sum_{i>0} \{C(\dot{A}_i \sim A)/C(E)\}i \end{aligned}$$

Since $C(E) > 0$, and $C(\sim A\dot{A}_i) > 0$ for some i, it follows that $D^+(A) > 0$.

$$\begin{aligned} D^+(A) &= D^+_A(A) && \text{by desiredness} \\ &= \sum_i C^+_A(\dot{A}_i)i && \text{by harmony} \\ &= \sum_i C(\dot{A}_i | A [A \rightarrow \dot{A}_0])i \\ &= \sum_i C(\dot{A}_i | A \dot{A}_0)i \\ &= \sum_{i=0} C(\dot{A}_i | A \dot{A}_0)i + \sum_{i>0} C(\dot{A}_i | A \dot{A}_0)i \end{aligned}$$

Now: $C(\dot{A}_0 | A \dot{A}_0) = 0$, and $C(\dot{A}_i | A \dot{A}_0) = 0$ if $i > 0$. Hence: $D^+(A) = 0$.⁶

⁶This horn of the contradiction can also be derived directly from the partition principle using the formula which in section 8 I call *Basic Harmony*: Where $D(A|B)$ is the desiredness of A given B, for the harmonious agent we clearly should have $D(A|\dot{A}_i) = i$.

$$D^+(A) = \sum_i C^+(\dot{A}_i | A) D^+(\dot{A}_i A) \quad \text{by the partition principle}$$

(1) and (2) are in contradiction. Our journey appears to have been in vain.

7 CHANGING VALUES

In my 1994 I claimed that the general version of harmony was immune to the style of *reductio* presented in Lewis’s first paper. Briefly, I argued that when we are dealing with propositions the truth values of which change over time (i.e. the ranges of which are classes of world-times rather than classes of worlds) conditionalization is evidently problematic. Consider a changing proposition which is not about value at all, say *it is snowing* . On Monday I am standing outside with my eyes closed, gently warming my nose in the Colorado sun, and (in the light of my total confidence in *my nose is warm*) my credence in *it is snowing* is 0. On Tuesday I am still standing outside, eyes closed, but there are flakes accumulating on my nose, and in the light of *my nose is frozen* my degree of belief in *it is snowing* rockets to 1. By Wednesday, my nose has thawed out in that reliable Southwestern sun, and given my renewed and complete confidence that *my nose is warm*, my belief in *it is snowing* drops back to 0. Clearly my Tuesday credence is not obtained from my Monday credence by updating on the new information. The claim that my Tuesday credence in *it is snowing* should equal my Monday credence updated by the new information that *my nose is frozen* is crazy! If I were to update in that way then my confidence in *it is snowing* would remain at zero. To apply conditionalization straightforwardly we need to traffic solely in dated timeless propositions.

Propositions about the goodness of a state of affairs can also change their truth values. To apply conditionalization we have to traffic in unchanging dates propositions. It turned out that on all the plausible reworkings of Lewis’s original argument, so that all changing propositions are replaced by dated propositions, at least one step in the argument failed.

How does this analysis fare on the current simplified proof? At first blush it does look applicable. Redescribe Frederic’s situation from the objective point of view. Let us suppose, for definiteness, that value is an objective analogue of desiredness. The value of a

$$\begin{aligned}
 &= \sum_i C^+(\hat{A}_i|A)D^+(\hat{A}_i|A) && \text{(since } D(A|B) = D(AB)) \\
 &= \sum_i C(\hat{A}_i|A[A\hat{A}_0])i && \text{(Basic Harmony)} \\
 &= \sum_i C(\hat{A}_i|A\hat{A}_0)i
 \end{aligned}$$

Since $C(\hat{A}_0|A\hat{A}_0)=1$, $D^+(A) = 0$.

state is, let us suppose, the chance-weighted average of the objective value of possible outcomes conditional upon the state in question obtaining. So assuming that the pirates will toss a fair coin to decide whether or not to release Frederic, prior to the coin-tossing the value of an early marriage will be $1/2$, as will the value of a late marriage. After the coin-tossing, the value of an early marriage will either zoom to 1 or plummet to 0. In this set-up the value of an early marriage is a changing magnitude, and so propositions like \mathring{A}_0 and \mathring{A}_1 also change their truth-value over time. So it seems after all that my earlier analysis might apply.

Too swift. We can purge the example of changing propositions, and whatever problems they create for updating by conditionalization. One way to do this is to suppose that the pirates have already made their decision about an early release, it is now settled, although Frederic has no idea what that decision is. All he knows is that their decision, and with it the value of an early marriage, has already been settled by the toss of a fair coin, and is now fixed and unchanging. Frederic may divide his credence equally between the proposition that the value of an early marriage is 1, and the proposition that its value is 0. He still has to opt for an early marriage or a late one, but he cannot make up his mind, since his best estimate of the value of both options is that they are of the same middling value, $1/2$. The initial assumption (#) of the proof is satisfied since Frederic has no idea either what he will do or what the pirates have decided.

Now imagine Frederic receives the following information E: *if you choose an early marriage then the pirates decided not to release you from the contract*. This, of course, is tantamount to the proposition *if you choose an early marriage then an early marriage is of zero value*: $[A \emptyset \mathring{A}_0]$. He dutifully updates on this information to C^+ . By the theorem we know he cannot satisfy Harmony. And we can now see why. By assumption he still gives some credence to the proposition that the pirates have chosen to release him. So he still gives some credence to the proposition that an early marriage is a good thing. His estimate of the value of A is thus non-zero (first horn). But given his new information E, the desiredness of A will be 0 (second horn).

8 NEWCOMB RETURNS

The example clearly has a whiff of Newcomb about it. In conjunction with his new information, E, the proposition that he chooses an early marriage, informs him about the situation which settles the desirability of that very choice. Choosing (and thereby learning) A gives him the information that the pirates have decided not to release him, and that makes A bad. $C^+(\mathring{A}_0|A) < C^+(\mathring{A}_0)$. It is precisely when one's choices alter the probabilities of the

range of those possible settled conditions which make the action more or less desirable, that Newcomb problems arise.

Here is one formulation of causal decision theory. Take the partition K_1, \dots, K_n which (at a given moment) yields the mutually exclusive and jointly exhaustive set of possibilities for the past and laws, together with whatever is settled by the past and laws. Whereas standard subjective decision theory tells us that for any partition whatsoever:

$$D(A) = \sum_j C(K_j|A)D(AK_j)$$

causal decision theory tells us that the choice-worthiness (W) of A is given rather by:

$$W(A) = \sum_j C(K_j)D(AK_j).^7$$

D and W may, of course, come apart when for one or more j , $C(K_j|A) \neq C(K_j)$.

Let us make two further suppositions. First, suppose that each member K_j of the partition entails a particular proposition about the *value* (note: not just desiredness) of A : $Val(A)=k_j$. For example, we could take the (objective) value of A to be the *chance* weighted average of the value of the feasible futures compatible with A . Since the past and laws jointly settle all the chances, each K_j will entail one and only one of the \hat{A}_i . Second, consider what the realist should say about the desiredness of A *given the value of A*. For a harmonious agent, *the desiredness should equal the value given*. Call this *basic harmony*:

Basic Harmony

The desiredness of A , given the value of A , is the given value itself:

$$D(A|\hat{A}_i) = i.$$

Since the K_j s form a partition which is finer than the partition formed by the \hat{A}_i s, each \hat{A}_i is tantamount to a disjunction, $K_1^{(i)} \vee \dots \vee K_n^{(i)}$, and for the harmonious agent $D(A|K_h^{(i)}) =$

$D(A|\hat{A}_i)$, for $1 \leq h \leq n$.

Hence:

$$\begin{aligned} Est(V(A)) &= \sum_i C(\hat{A}_i)i \\ &= \sum_i C(\hat{A}_i)D(A|\hat{A}_i) \\ &= \sum_i C(K_1^{(i)} \vee \dots \vee K_n^{(i)})D(A|\hat{A}_i) \end{aligned}$$

⁷See Lewis 1986, p. 312ff.

$$\begin{aligned}
&= \sum_i \sum_h C(K_h^{(i)}) D(A|K_h^{(i)}) \\
&= \sum_j C(K_j) D(A|K_j) \\
&= W(A).
\end{aligned}$$

So, under these assumptions, what the *BAD* paradox reveals is just that $D(A)$ and $W(A)$ (which, assuming basic harmony, is $\text{Est}(V(A))$) can come apart—namely, when $C(K_j) \neq C(K_j|A)$. The *BAD* paradox is thus a Newcomb problem in disguise.

Is Newcomb a problem for the ideal of Harmony? Does Newcomb show that one cannot or should not desire things just to the extent that one believes them to be valuable?

Suppose we think that causal decision theory gives the intuitively right result for the desirability of actions. Rational desire should, then, be aligned with choice-worthiness (W), rather than what we have been calling desiredness (D). The rationally motivated agent desires things precisely to the extent that they are choice-worthy. But then, the rationally motivated agent must satisfy harmony: she desires things precisely to the extent that she believes them to be valuable. That's not too bad after all.

9 THE SUBJECTIVIST'S REVENGE?

Return to Frederic in his original dilemma. He knows the pirates have chosen whether to release him or not, but he is at a loss to know what they have chosen. He divides his credence equally between the four possible outcomes ($A\dot{A}_1, \sim A\dot{A}_1, A\sim\dot{A}_0, \sim A\sim\dot{A}_0$). Then he learns $E = [A \rightarrow \dot{A}_0]$ from his “authoritative source”—perhaps a friendly pirate who was in on the decision making.

Forget about harmony *et al* for the moment and just concentrate on the relative merits of Frederic's two choices. It is clear what it is rational for Frederic, in his epistemic situation, to do, *viz.* delay the marriage. Compare A and $\sim A$ for desiredness after E comes in and we get: $D^+(A) = 0 < D^+(\sim A) = 1/2$. Compare them for estimated value, or choice-worthiness, and we have: $W^+(A) = 1/3 < W^+(\sim A) = 2/3$. Thus both D^+ and W^+ recommend deferring marriage.

So far, then, we have a result which is problematic for the original explication of harmony or desire-as-belief, but one which is not in itself paradoxical. Both W and D apparently give the intuitively right judgement on what it is rational for Frederic to do.

Can we arrange it that W and D make different recommendations in this situation? That shouldn't be too hard, provided we rig up the initial credences appropriately. Let the four outcomes have the following initial credences:

P	C(P)
$A\dot{A}_1$	0.15
$\sim A\dot{A}_1$	0.45
$A\dot{A}_0$	0.10
$\sim A\dot{A}_0$	0.30

Don't ask where Frederic got these credences from. Let's just assume they have been foisted on him. Then when Frederic learns $[A \rightarrow \dot{A}_0]$ he apparently must revise by conditionalization to:

P	$C^+(P)$
$A\dot{A}_1$	0
$\sim A\dot{A}_1$	45/85
$A\dot{A}_0$	10/85
$\sim A\dot{A}_0$	30/85

Now he finds that his estimate of the value of A is higher than his estimate of the value of $\sim A$, while $D(A)$ remains at 0 and $D(\sim A)$ is positive. Not only do W and D fall apart here, but it is clear which measure gives the right ordering of A and $\sim A$: namely, *desiredness!*

What is going on here? When Frederic learns $[A \rightarrow \dot{A}_0]$ from the friendly pirate is that all he learns? Let us suppose that the pirate has sworn an oath not to tell Frederic what they decided. Has he broken his oath? Technically, no. The information he explicitly imparted was a proposition which leaves both of the pirate options open. And yet having been told this information Frederic, knowing that the decision has been made and cannot be undone by what he chooses, surely knows what their decision was. Since he cannot alter the past, but he can choose between A and $\sim A$, he knows that if he chooses A the result will be the conjunction of A and whatever it is the pirates decided. Since he knows that if he chooses A the result will be $A\dot{A}_0$, he thereby knows the pirate's decision. So putting his informer's message together with what he knows already, he really learns something stronger: *viz* \dot{A}_0 . So his actual credences should look like this:

P	C(P)
---	------

$\dot{A}\dot{A}_1$	0
$\sim\dot{A}\dot{A}_1$	0
$\dot{A}\dot{A}_0$	1/4
$\sim\dot{A}\dot{A}_0$	3/4

And now Frederic's estimate of the value of A is 0 and his estimate of the value of $\sim A$ is 1, giving us the right result after all.

10 **BAD ARGUMENTS ARE GOOD NEWS!**

Desiring things just to the degree one believes them to be valuable is an attainable ideal, one which does not conflict with decision theory properly understood. The standard arguments against the possibility of *desire-as-belief* thus do no damage to the realist's ideal of maintaining harmony between one's desires and one's beliefs about the good. Somewhat surprisingly, those arguments actually turn out to be variants of Newcomb's problem, and rightly understood they reinforce a version of causal decision theory. Thus these particular *BAD* arguments against value realism turn out to be *GOOD NEWS* for casual decision theory.

References

Byrne, A. and Hajek, A. “David Hume, David Lewis, and Decision Theory” *Mind* 106, pp. 411-27.

Oddie, G. “Harmony, Purity, Truth” *Mind*, 103: 452-72 (1994).

Lewis, D. “Desire as Belief” , *Mind* 97, pp. 323-32 (1988).

Lewis, D “Desire as Belief II”, *Mind* 103, pp. 303-13.