David Hume, David Lewis, and Decision Theory

ALEX BYRNE AND ALAN HÁJEK

David Lewis claims that a simple sort of anti-Humeanism—that the rational agent desires something to the extent he believes it to be good—can be given a decision-theoretic formulation, which Lewis calls "Desire as Belief" (DAB). Given the (widely held) assumption that Jeffrey conditionalising is a rationally permissible way to change one's mind in the face of new evidence, Lewis proves that DAB leads to absurdity. Thus, according to Lewis, the simple form of anti-Humeanism stands refuted.

In this paper we investigate whether Lewis's case against DAB can be strengthened by examining how it fares under rival versions of decision theory, including other conceptions of rational ways to change one's mind. We prove a stronger version of Lewis's result in "Desire as Belief II". We then argue that the anti-Humean may escape Lewis's argument either by adopting a version of causal decision theory, or by claiming that the refutation only applies to hyper-idealised rational agents, or by denying that the decision-theoretic framework has the expressive capacity to formulate anti-Humeanism.

1. Introduction

Does rationality demand that one have certain desires? Is it impossible to have certain beliefs without also having certain desires? Even if it is not, does rationality demand that certain desires and beliefs go together? Humeans, like David Hume and David Lewis, answer "no" to such questions. Lewis (1988, 1996) has applied the resources of decision theory in support of the Humean cause. More exactly, Lewis claims that a simple sort of anti-Humeanism—that the rational agent desires something to the extent he believes it to be good—can be given a decision-theoretic formulation, which Lewis calls "Desire as Belief" (DAB). Given the (widely held) assumption that Jeffrey conditionalising is a rationally permissible way to change one's mind in the face of new evidence, Lewis (1988) proves that DAB leads to absurdity. (Lewis 1996, and Arló Costa, Collins and Levi 1995 supply simpler proofs for the special case of conditionalising.[1]) Thus, according to Lewis, the simple form of anti-Humeanism stands refuted.

[1] For a refutation of the counterpart of DAB in non-quantitative decision theory, see Collins (1988).

 © Oxford University Press 1997

Huw Price (1989) claims to provide another decision-theoretic formulation of an anti-Humean thesis close to Lewis's original, and shows that it escapes Lewis's 1988 result. Lewis (1996) proves that Price's formulation (dubbed by Lewis "Desire as Conditional Belief" (DACB)) is equivalent to one Lewis calls "Desire by Necessity" (DBN). Lewis argues, contrary to Price's advertisement, that this equivalence shows DACB not to be a version of "the anti-Humean idea that desires and beliefs are necessarily aligned" (p. 313). Rather, according to Lewis, DACB is a formulation of another anti-Humean idea: that rationality demands that one have certain desires.

In this paper we shall investigate whether Lewis's case against DAB can be strengthened by examining how it fares under rival versions of decision theory, including other conceptions of rational ways to change one's mind. The answer is both "yes" and "no". We conclude by considering some anti-Humean replies.

2. *Preliminaries*

At any time, the mental state of an (idealised) rational agent determines a pair of functions $<C, V>$. C is the agent's credence function, from propositions to the unit interval; V is the agent's value function, from propositions to real numbers.[2] Propositions are sets of possible worlds; of special interest are the necessary proposition I (true at all worlds), the impossible proposition $\emptyset$ (true at none), and various point-propositions (true at exactly one world).[3] The credence function C is assumed to be a probability function (and so to obey the usual axioms). As the set of propositions forms a boolean algebra, we can speak of the conjunction XY (intersection), disjunction $X \vee Y$ (union), and negation $\sim X$ (complement) of propositions X and Y. Further, X implies Y iff X is a subset of Y ($X \subseteq Y$). The value function V is standardly taken to obey the following rule of additivity: for all propositions A such that $C(A) \neq 0$,

$$V(A) = \Sigma_i V(W_i).\ C(W_i/A)$$

where the W_i are the point-propositions, and where "$C(W_i/A)$" is defined as the ratio $C(W_iA)/C(A)$.

The values an agent assigns to point-propositions—his point-values—are assumed not to c2 2hange as he learns. When an agent learns A, his credence function changes, and so—given the previous assumption—the change in his value function is determined by the rule of

[2] Strictly speaking there is no such thing as *the* value function of an agent: the value function is only unique up to fractional linear transformations.

[3] For simplicity we shall assume that there are countably many worlds.

additivity. Say that a *rule of updating* is a way an agent may change his total state of opinion when his total evidence is A, with the proviso that his opinion must at least obey the probability calculus. So, a rule of updating specifies, for all propositions A and credence functions C, a new credence function C_A, such that $C_A(A) = 1$.[4] Conditionalising is one such rule of updating, where, if $C(A) > 0$, $C_A(X) = C(X/A)$. As required, $C(A/A) = 1$.

Although conditionalising is the most commonly favoured rule to model rational change of mind, others have been suggested. Example: *imaging* on A is the rule that moves the probability the agent's initial credence function assigns to world w (equivalently, to each point-proposition W) to the world nearest w at which A is true, where the similarity measure between worlds is borrowed from Stalnaker's (1968) possible worlds analysis of counterfactuals. Lewis (1976) shows that, if C_A is the result of imaging on A applied to an initial credence function C, then $C_A(X) = C(A > X)$, where the function $>$ is the Stalnaker conditional ("$A > X$" is true at those worlds w where the A-world closest to w is an X-world). (Dropping the requirement that the probability moved from w is concentrated on just one world gives (various versions of) *blurred* or *general* imaging.[5])

In some intuitive sense of "value", $V(A)$ is supposed to be a measure of the value the agent attaches to A. With only this loose constraint in mind, let us ask how we might determine the right rule of additivity for value (not assuming, then, that it must be the standard rule given above). The value of a proposition A (for an agent) plausibly ought to be a weighted sum of the values of the different ways W_i it could come true. But how are the weights to be calculated? Perhaps the weights should be determined solely by the credences the agent *does* give to each of these ways, the credence he *does* give to A, and perhaps to other propositions. Alternatively, perhaps the weights should be the credences the agent *would* give to each of these ways, were A to be his total evidence. If the first answer is correct, then taking the weights to be the familiar ratios is the obvious option. However, if the second answer is correct, then the additivity rule becomes:

$$V(A) = \Sigma_i V(W_i).\ C_A(W_i).$$

If the agent updates by conditionalising then, of course, both answers yield the same additivity rule. But we will have divergent conceptions of the rule if the agent employs some other method of updating. Say that V *is ratio-weighted* iff the additivity rule is given by the usual ratio formula, and that V *is updated* iff the rule is given by the formula immediately

[4] Jeffrey conditionalising is not a rule of updating in our sense, for it allows the agent to change his opinion simply by becoming more confident, or less confident, in A. See Lewis (1988) and Jeffrey (1965).

[5] See Gärdenfors (1982) and Lewis (1981).

above. V will be both ratio-weighted and updated iff the additivity rule is the usual formula and the agent updates by conditionalising.

There is another possible view about the right rule of additivity: to calculate $V(A)$, sum over the values of the different ways W_i that A could come true, and weight the value of each way W_i by the agent's credence in whether making A true would make W_i true. This gives a version of *causal* decision theory (see Lewis 1981). With an appropriate similarity measure, chosen to rule out a "back-tracking" interpretation of counterfactuals,[6] the rule of additivity may then be taken to be

$$V(A) = \Sigma_i V(W_i).\ C(A > W_i).$$

Say that V *is causalised* in this case. V will be both causalised and updated iff the additivity rule is this formula and the agent updates by imaging (with the measure chosen to rule out back-trackers). If V is causalised and the agent updates by conditionalising (for example), then V is neither ratio-weighted nor updated. In fact, updating by conditionalising and taking V to be causalised is an attractive combination: there are powerful arguments for each part of the package.[7] And, if V is supposed to be the quantity that the rational agent will act so as to maximise, this combination is, near enough, the version of decision theory preferred by Lewis himself. Lewis does not think, however, that causal decision theory is suitable for the formulation of an anti-Humean thesis (1996, p. 304). To avoid complicating matters unduly, we shall start by restricting attention only to cases where V is either ratio-weighted or updated. This restriction will be briefly lifted in §3, Part C. We shall address Lewis's objection to using causal decision theory for the formulation of anti-Humean theses in §4.

When we speak of *all rational* $<C, V>$ (credence-value) pairs (sometimes we shall drop "rational") we mean: all $<C, V>$ pairs that correspond to some possible (idealised) rational agent. We shall assume that any credence function is the credence function of some rational agent.[8] When we speak of *all* (rational) V, we mean: all V such that V is the second member of some rational credence-value pair. We shall further suppose that, for all A, the rational agent may, when updating on A, change his credence func-

[6] Sometimes the contextually supplied similarity measure makes it true to say that if the present had been different in so-and-so ways the past would have been different in such-and-such ways—these are cases where a counterfactual is interpreted in a "back-tracking" sense. We need to rule these out if counterfactuals are supposed to express some hypothetical causal connection between the consequent and the antecedent. On back-tracking, see Lewis (1979).

[7] Namely, arguments for conditionalising based on diachronic Dutch books (see, e.g., Teller 1973), and arguments for causal decision theory based on Newcomb's problem (see, e.g., Lewis 1981).

[8] This assumption could be questioned and, for what follows, could be considerably weakened. Partly for this reason, we will not question it.

tion from C to a certain new credence function C_A, with his value function V changing (in accordance with the chosen rule of additivity) to V_A. So if $<C, V>$ is a rational credence-value pair, then, for all A, so is $<C_A, V_A>$. We shall assume nothing more about all rational $<C, V>$ pairs. The following three decision-theoretic formulations of putative anti-Humean theses will impose various constraints on this set.

> *Desire as Belief* (DAB)
>
> There is a function ° (a "halo" function) such that for all (non-empty) propositions A and for all rational $<C, V>$:
>
> $V(A) = C(A°)$.

Anti-Humean gloss: "[n]ecessarily, and regardless of one's credence distribution, one must desire A exactly to the extent that one believes it to be good" (Lewis 1996, p. 308).

Note that DAB—like the following two theses—will come in various versions, depending on how we specify the rule of updating and the rule of additivity for V.

> *Desire as Updated Belief* (DAUB)
>
> There is a function ° such that for all propositions A and for all rational $<C, V>$ such that $C(A) \neq 0$:
>
> $V(A) = C_A(A°)$.

Where updating goes by conditionalising, DAUB becomes Price's:

> *Desire as Conditional Belief* (DACB):
>
> $V(A) = C(A°/A)$.

Anti-Humean gloss: "DACB equates [degrees of desire] to conditional credences" (Lewis 1996, p. 309).

Now restrict the value functions to those V such that, for all point-propositions W, $V(W)$ is either 0 or 1. The final anti-Humean formulation is:

> *(Generalised) Desire by Necessity* (GDBN)
>
> There is a unique proposition—call it "G"—such that, for all V, W, $V(W) = 1$ if W implies G, $V(W) = 0$ otherwise.

Anti-Humean gloss: the rational agent desires The Good (cf. Lewis 1996).

Applying the chosen rule of additivity for value, it follows from GDBN that, for all propositions A and for all $<C, V>$ such that $C(A) \neq 0$:

> $V(A) = C_A(G)$ (if V is updated),
>
> $V(A) = C(G/A)$ (if V is ratio-weighted).

In the latter case, GDBN becomes Lewis's *Desire by Necessity* (DBN).

3. Some results

First, in Part A, we shall show that DAUB is equivalent to GDBN, assuming that V is updated. Second, in Part B, we shall show that DAB leads to absurdity, assuming (a) that V is ratio-weighted, and (b) that the method of updating obeys a very weak constraint. Third, in Part C, we shall briefly investigate how DAB fares if V is updated, and finally consider the case where V is causalised and updating goes by conditionalising.

Part A

Lemma 1

DAUB implies that, for any point-proposition W, either, for all V, $V(W) = 1$, or, for all V, $V(W) = 0$.

Proof

By DAUB, for any W and $<C, V>$, $V(W) = C_W(W^\circ)$. $C_W(W^\circ)$ is either 1 (if $W \subseteq W^\circ$) or 0 (otherwise). If it's 1, then, for all V, $V(W) = 1$; if it's 0, then, for all V, $V(W) = 0$.

Lemma 2

DAUB implies GDBN, and GDBN (assuming that V is updated) implies DAUB.

Proof

(DAUB implies GDBN.) By lemma 1, DAUB implies that all value functions agree on the point-propositions, assigning them values of either 1 or 0. So we may let $G =_{df.} \cup_i W_i$ such that, for all V, $V(W) = 1$. (GDBN implies DAUB, assuming that V is updated.) For all A, let $A^\circ = G$.

An immediate corollary of lemma 2 is the equivalence of DACB and DBN (assuming that V is updated). (For a different proof of this equivalence, see Lewis 1996.)

Part B

Lemma 3

DAB implies that, for any point-proposition W, either, for all V, $V(W) = 1$, or, for all V, $V(W) = 0$.

Proof

By DAB, for any W and $<C, V>$, $V(W) = C(W^\circ)$. Since $V(W)$ is unchanged by any updating, so is $C(W^\circ)$. Hence W° is either I or $\emptyset$. If it's I, then for all V, $V(W) = 1$. If it's $\emptyset$, then, for all V, $V(W) = 0$.

For the remainder of Part B we shall assume that V is ratio-weighted.

Lemma 4

DAB implies DBN.

Proof

By lemma 3, and the argument of the first part of the proof of lemma 2.

Lemma 5

DAB implies: for all A, C, if $C(A) \neq 0$, $C(A^\circ) = C(G/A)$.[9]

Proof

By lemma 4.

Let *Moderation* be the following constraint on updating: for all propositions X, Y, and credence functions C, if X implies Y, and $C(X) > 0$, then $C_Y(X) > 0$.[10] That is, if the agent gives some credence to X, and updates on something that X implies, he still gives some credence to X. That it is rationally *permissible* (not necessarily rationally *required*) to update by a rule that obeys Moderation is—modulo the permissibility of *updating* by some rule—more a desideratum than an assumption. Moderation holds of, inter alia, conditionalising, imaging (assuming "centering": any A-world w is the closest A-world to w), and blurred imaging (assuming "weak centering": any A-world w is among the closest A-worlds to w).

Theorem

Assuming that the rule of updating obeys Moderation, DAB implies that, for all A such that $\emptyset \neq A \neq I$, either $A^\circ = \emptyset$ or $A^\circ = I$ (i.e. for all contingent A, A° is non-contingent).

[9] This implication of DAB equates certain unconditional probabilities with certain conditional probabilities. It is reminiscent of a notorious hypothesis linking the probabilities of conditionals and conditional probabilities (challenged by Lewis 1976):

there is a function $\rightarrow$ such that, for all A, B, C, if $C(A) \neq 0$, $C(A \rightarrow B) = C(B/A)$.

There are results that show that this hypothesis leads to triviality, without any further assumption about the proposition $A \rightarrow B$. In particular, they work even if $A \rightarrow B$ is a function solely of A, which we might represent as "$A^{\copyright}$". Furthermore, some of them have force even if the domain of propositions over which the equation is supposed to hold is restricted (see especially Hájek 1994 and 1996)—as it might be, allowing B to be only a certain proposition G. Then we have:

there is a function $^{\copyright}$ such that, for all A, C, if $C(A) \neq 0$, $C(A^{\copyright}) = C(G/A)$.

Not surprisingly, then, similar methods can be deployed to show the triviality of DAB, once we have shown that it implies something of the same form.

[10] Moderation is first introduced in Hájek (1996).

Proof

Suppose not. Then possibly: the rule of updating obeys Moderation, DAB holds, and for some contingent A, A° is contingent. Now either A° implies AG, or else it is consistent with $A{\sim}G$, or else it is consistent with ${\sim}A$. That is, either (i) $A^\circ \subseteq AG$, or (ii) $A^\circ A{\sim}G \neq \emptyset$, or (iii) $A^\circ{\sim}A \neq \emptyset$.

By lemma 5: (*) for all C, if $C(A) \neq 0$, $C(A^\circ) = C(G/A)$.

Suppose $AG = \emptyset$. Then, by (*), for all C, if $C(A) \neq 0$, $C(A^\circ) = C(GA)/C(A) = C(\emptyset)/C(A) = 0$. Hence $A^\circ = \emptyset$, contradicting part of our initial assumption. So $AG \neq \emptyset$.

Case (i). Suppose $A^\circ \subseteq AG$. As $\emptyset \neq A \neq I$, and $AG \neq \emptyset$, we can choose some C such that $0 < C(A) < 1$, and $C(AG) > 0$. Now $C(A^\circ) \leq C(AG) < C(AG)/C(A)$. So $C(A^\circ) < C(G/A)$, contradicting (*). So not-(i).

Case (ii). Suppose $A^\circ A{\sim}G \neq \emptyset$. Choose a C such that $C(A^\circ A{\sim}G) > 0$. (Hence $C(A) > 0$.) By (*), and updating on $A^\circ A{\sim}G$ we have $C_{A^\circ A{\sim}G}(A^\circ) = C_{A^\circ A{\sim}G}(G/A)$. But $C_{A^\circ A{\sim}G}(A^\circ) = 1$, and $C_{A^\circ A{\sim}G}(G/A) = C_{A^\circ A{\sim}G}(GA)/C_{A^\circ A{\sim}G}(A) = 0$, contradiction. So not-(ii).

Now, if $A \subseteq G$, then by (*), for all C, if $C(A) \neq 0$, $C(A^\circ) = 1$, so $A^\circ = I$, contradicting the contingency of A°. So $A{\sim}G \neq \emptyset$. Hence ${\sim}(GA) \neq \emptyset$.

Case (iii). Suppose $A^\circ{\sim}A \neq \emptyset$. (Now $A{\sim}G \neq \emptyset$.) Choose some C such that $C(A^\circ{\sim}A), C(A{\sim}G) > 0$. (Hence $C(A) > 0, C({\sim}(GA)) > 0$.) By (*), and updating on ${\sim}(GA)$ we have $C_{{\sim}(GA)}(A^\circ) = C_{{\sim}(GA)}(GA)/C_{{\sim}(GA)}(A)$. (By Moderation, the fact that $A{\sim}G \subseteq {\sim}(GA)$, and the fact that $C(A{\sim}G) > 0$, we have $C_{{\sim}(GA)}(A{\sim}G) > 0$. Hence $C_{{\sim}(GA)}(A) > 0$, and so the ratio $C_{{\sim}(GA)}(GA)/C_{{\sim}(GA)}(A)$ is well-defined.) By Moderation, the fact that $A^\circ{\sim}A \subseteq {\sim}(GA)$, and the fact that $C(A^\circ{\sim}A) > 0$, we have $C_{{\sim}(GA)}(A^\circ{\sim}A) > 0$. Hence $C_{{\sim}(GA)}(A^\circ) > 0$. But $C_{{\sim}(GA)}(GA)/C_{{\sim}(GA)}(A) = 0$, contradicting $C_{{\sim}(GA)}(A^\circ) = C_{{\sim}(GA)}(GA)/C_{{\sim}(GA)}(A)$. So not-(iii).

Therefore, if updating is Moderate, it is impossible that DAB holds, and for some contingent A, A° is contingent. So, if updating is Moderate, DAB implies that, for all contingent A, A° is non-contingent. *QED.*

Thus, if updating is Moderate, DAB (recall we are assuming that V is ratio-weighted) implies that for *any* contingent proposition A, for all $\langle C, V\rangle$, either $V(A) = C(I) = 1$ or $V(A) = C(\emptyset) = 0$. That is, all rational agents attach the same value to A. In particular, a rational agent never changes his mind about the value he attaches to A, no matter what he learns. This is absurd.

Obviously, interpreting "A°" as "A is good" does not make the absurdity go away.[11] Therefore (given ratio-weighting and Moderation), *it is not*

[11] If anything, it makes matters worse. For numerous contingent propositions A, the proposition that A is good is contingent. Example: let A be the proposition that Lara is wealthy. A's goodness depends on contingent matters, of which a rational agent may or may not be aware: whether, for instance, Lara uses her money shopping at Frederick's of Hollywood or to promote racial harmony in Los Angeles.

a requirement of rationality that one value something to the degree one believes it to be good.[12] (And here, of course, we do *not* need to assume that V measures the agent's degree of desire.)

Part C

Suppose, finally, that V is updated, and that, for all A, if $C(A) = 1$, updating on A leaves the credence function unchanged. (This last supposition really is a desideratum.) It follows, given that V is updated, that updating on A leaves $V(A)$ unchanged, and so DAB implies DAUB, where there is a halo function that serves for both cases. Now DAUB and DAB (together with the fact that there is halo function that serves for both) imply:

> *(Generalised) Independence* (GIND)
>
> There is a function ° such that for all propositions A and for all C such that $C(A) \neq 0$:
>
> $$C(A^\circ) = C_A(A^\circ).$$

Where updating goes by conditionalising, GIND becomes Lewis's IND:

$$C(A^\circ) = C(A^\circ/A).$$

Hence DAB implies GIND. (This is simply a more general version of Lewis's proof (1996) that DAB, assuming conditionalising and ratio-weighting, implies DACB and IND.[13]) In the special case of conditionalising (i.e. GIND = IND), Lewis shows that GIND leads to absurdity. What about other ways of updating?

It is not difficult to replicate Lewis's result for a wide range of updating rules. Imaging, however, is a special case. As we remarked in §2, if C_A is the result of imaging an initial credence function C on A, then $C_A(X) = C(A > X)$, where the function > is the Stalnaker conditional. Therefore, if imaging is used both to update and to calculate V, then DAB implies that there is a function ° such that for all A, C, $(C(A) > 0)$:

$$C(A^\circ) = C(A > A^\circ).$$

[12] We need another assumption as well: that idealised rational agents obey (or, at least, may obey) decision theory. We will raise some doubts about this in §4 below.

[13] When Lewis says that "IND follows immediately from DAB and DACB", and "DAB follows from DACB and IND" (1996, p. 309), he is talking about the corresponding equations, all taken to be within the scope of the halo function quantifier "there is a function °". As we have elected to use these terms, it is not true that DAB and DACB follow from IND, nor that DAB follows from DACB and IND.

This does not imply (absurdly) that A° is non-contingent.[14] One might think, though, that it is still problematic. In some cases, it seems, it can be perfectly rational to give different credences to A° (read as "A is good") and $A > A^\circ$ (read as "if A had been true, A would have been good"). Here is an example. Suppose you believe that Dr. Feelgood has your best interests at heart, and prescribes drug alpha instead of drug beta. You might, therefore, give low credence to the proposition that taking drug beta is good. But you might also give high credence to the proposition that if you had taken drug beta, taking it would have been good. For if you had taken beta, that would have been because Feelgood prescribed it, in which case, given the good doctor's benevolence, taking beta would have been good. If this is your state of mind, you are not thereby irrational.

No problem, however. The above reasoning assumes that the counterfactual is taken in a "back-tracking" sense, which allows us to infer from hypothesised different effects back to different causes and their consequences. Arguably, we standardly use a similarity measure to evaluate counterfactuals that rules out such back-tracking (Lewis 1979). Even if this is incorrect, the counterfactual proposition $X > Y$ must be taken in a non-backtracking sense, as we are using it to express some (loose sort of) causal dependency of Y's truth on X's truth. Hence the reasoning fails. And since it appears (to us) that any such alleged counterexample will involve back-tracking, taking V to be updated and updating by imaging (with the appropriate measure) leaves DAB unmarked.

That completes our discussion of the cases where V is either ratio-weighted or updated. We now relax this assumption, and consider the case where V is causalised and updating goes by conditionalising (as we noted in §2 this is, near enough, Lewis's preferred formulation of causal decision theory). So the additivity rule is exactly the same as in the case mentioned at the end of the previous paragraph, but the updating rule is conditionalising, not imaging. It does not follow from these assumptions that the rational agent gives the same credence to A° and $A > A^\circ$ (in any case, as we argued, such an implication would not be problematic). More importantly, none of the Lewis-style anti-DAB proofs can be replicated against this combination of updating and value-additivity. As far as we can determine, this version of causal decision theory (like the case immediately above where updating goes by imaging) coexists peaceably with DAB.[15]

So the results are a mixed bag. If V is ratio-weighted, and the method of updating obeys Moderation, then DAB is absurd. Similarly if V is

[14] Since the equation holds for *all* C, $(C(A) > 0)$, $A^\circ = A > A^\circ$. And (as Robert Stalnaker pointed out to us) if we defined A° as $A > G$, then it would follow from this fact alone that $A^\circ = A > A^\circ$, since $X > (X > Y)$ is logically equivalent to $X > Y$.

[15] John Collins makes the same point in his PhD thesis (1991).

updated and updating goes by conditionalising. If V is causalised, then whether updating goes by imaging (in which case V is also updated) or by conditionalising, DAB does not appear to run into problems.

4. Anti-Humean responses

These are of two kinds. The first kind accepts that anti-Humean theses can be formulated in decision theory with its usual framework assumptions that we set out in §2. In particular, the first kind of response concedes (a) that the decision-theoretic rational agent is the rational agent a sensible anti-Humean has in mind, and (b) that $V(A)$ can be thought of as specifying the agent's degree of desire in A. The second kind questions these assumptions.

First, let us consider what has been shown about Price's version (and its variants) of anti-Humeanism. (The following is hardly a *response* to Lewis's arguments, instead more an elaboration of what he says himself.) Part A of the previous section showed that DAUB is equivalent to GDBN, assuming that V is updated. (More exactly: DAUB implies GDBN, and GDBN and the assumption that V is updated imply DAUB.) GDBN simply says that, necessarily, any rational agent desires The Good. (Remember we are not now questioning that V specifies the agent's degree of desire.) GDBN is perfectly consistent, and is certainly an anti-Humean thesis (cf. "'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger" (Hume 1739–40, II, iii)). But plainly GDBN does not express the anti-Humean thought that there is a rationally necessary connection between an agent's contingent desires and his beliefs. DAUB, however, might appear to do just this. To take the case where updating goes by conditionalising, DAUB becomes: $V(A) = C(A^{\circ}/A)$ (i.e. DACB), which *does* express a necessary connection between desire and belief. But DAUB is derivable from GDBN with the decision-theoretic framework assumption that V is updated. Thus this necessary connection is *no more* than the one supplied by a certain conception of decision theory, and so is too weak to count as a specific anti-Humean doctrine. This is in essence the point Lewis is making when he says that DACB (i.e., DAUB in the case where updating goes by conditionalising) "is not, despite superficial appearances, a theory of contingent desire necessarily aligned with belief" (1996, p. 311). Nonetheless, DAUB (and GDBN) at least offer the "rich reward" of objective ethics (Lewis 1996, p. 307), and for all decision theory says on the matter, are there for the taking.

Let us now turn to versions of anti-Humeanism that are, unlike DAUB and GDBN, theories of "contingent desire necessarily aligned with

belief". Given the conclusion at the end of Part C of the previous section, an obvious anti-Humean move is to causalise *V*, either by arguing that *V* should be updated and that updating should go by imaging, or by arguing, in the usual way, that Newcomb problems show the need for causal decision theory. Lewis does briefly speak to an opponent who takes the second line. Suppose that a Newcomb problem arises: you believe that bringing about *A* would lead to some small good, but doing so would supply convincing evidence that a disaster over which you have no influence will befall you. If *V* is ratio-weighted, then (since we have a Newcomb problem) the *V*-maximal option will be $\sim A$. Consider the action of making *A* the case. Lewis says:

> Should you perform that action?—Yes; your destiny is not a consideration, since that is outside your control. Do you desire to perform it?—No; you want good news, not bad. (p. 304)

According to Lewis and other causal decision theorists, the quantity that the rational agent will act so as to maximise is not *V* ratio-weighted (let's simply call this "*V*"), but rather *V* causalised (let's call this "*U*").[16] However, the quantity one ought to identify with the agent's *degree of desire* is, on Lewis's view, *V*, not *U*. Why? Because, supposedly, the rational agent, even though he *acts to bring about A*, will *desire* $\sim A$ (the *V*-maximal option) more than he desires *A* (the *U*-maximal option).

But we do not find this compelling. Why think that, in a Newcomb case, the rational agent will *not* bring about his heart's desire? Of course, we can see how someone might intuit that taking the *U*-maximal option is the *rational thing* to do. And we can see how someone might intuit that the rational agent *most desires* to bring about the *V*-maximal option. What we *don't* see is why the *conjunction* of these is at all intuitive.[17] And so we

[16] Not quite: different causal decision theorists will differ—for present purposes, unimportantly—on exactly how to formulate the quantity that the rational agent will act so as to maximise. In taking this to be *V* causalised, we are adopting Allan Gibbard and William Harper's (1978) version of causal decision theory (a development of an idea due to Robert Stalnaker); see also Lewis (1981).

[17] Suppose that the Newcomb problem is this. You value watching horror films, as opposed to reading philosophy. But you are sure that a predisposition towards serial killing tends to cause a fascination with things horrific. If you discover that you are watching horror films, then this is evidence that you have murderous inclinations, which you naturally value not at all. Being a causal decision theorist, you go right ahead and rent *Creepers* from the video store (the *U*-maximal option), as opposed to curling up with the latest issue of *Mind* (the *V*-maximal option). Let's grant that you do not welcome the bad news your action brings. But does that mean that you desire to read *Mind* more than you desire to see *Creepers*? No: that makes your renting the video wholly mysterious. We can explain your sense of dissatisfaction without supposing that you most wanted to read *Mind*. You hoped that you were not the kind of person who would choose *Creepers* over *Mind*, and things did not turn out that way. Hence your dissatisfaction.

don't see that formulating DAB using V causalised is at all problematic (modulo the assumption about the expressive capacity of decision theory).

Another possible anti-Humean response is this. It is built into DAB that A° is unique (because it is a function solely of A). The anti-DAB proofs depend on this fact. So, if we gloss "A°" as "A is good", this means that for any proposition A, there is a unique proposition that A is good. But this has been denied: the proposition expressed by "A is good" might be held to vary systematically with the (actual or idealised) psychological states of the speaker or his community, a view we may call *subjectivist*. In fact, Lewis himself holds a subjectivist account: "to be a value—to be good, near enough—means to be that which we are disposed, under ideal conditions, to desire to desire" (1989, p. 116).

This response, then, concedes that DAB formulates *a* version of anti-Humeanism, but claims that certain subjectivist theories are (a) more plausible and (b) supply the required necessary alignment of contingent desire with belief. However, whether any version of subjectivism can satisfy both these requirements is open to dispute (and disputed by Lewis).[18]

These are what we take to be the two most promising anti-Humean responses, supposing that the decision-theoretic framework has the resources to formulate anti-Humeanism. But we will now question this supposition, in two ways. First, we shall argue that the rational agent of decision theory is, relative to the rational agent of a sensible anti-Humean, over-idealised. Then we shall give two reasons why $V(A)$ cannot be straightforwardly identified with the agent's strength of desire in A.

For all point-propositions W, and value functions V, $V(W)$ is supposed to remain constant no matter what the agent learns. Call this *constancy*; it is a crucial assumption for all Lewis-style anti-DAB proofs. Constancy is, in practice, a useful simplification. When decision theory is used to model problems of rational choice (for example, whether to go to a party, the pub, or to see Lara) it is convenient to assume (a) that these are maximally specific possibilities, and (b) that the agent is not going to change his preferences between these alternatives. What we now need to examine is whether there is more to constancy than this.

We have been supposing that (idealised) rational agents conform to the decision-theoretic framework (with the assumption of constancy) set out in §2: in other words, that standard decision theory is a norm of rationality. But really all the Humean needs to suppose is that constancy is merely rationally *permissible*, not rationally *required*. For if rationality *permits* one to preserve the values of point-propositions no matter what one learns, then (under certain other assumptions we have discussed) DAB leads to absurdity. Now since obeying a *requirement* of rationality,

[18] This is explored further in Hájek and Pettit (1996).

together with conforming to something that is rationally *permissible*, can hardly lead to absurdity, if constancy is permissible the desire-as-belief thesis cannot be a requirement of rationality.

But, offhand, there would appear to be no difficulty for the anti-Humean here. First, to suppose that rational agents have point-values at all is already to idealise in the extreme. Real-life examples of an agent with point-values are hard to come by. Perhaps this is one: a Nietzschean may value that *actual* history endlessly repeat itself, exactly as it happened, starting at the end of the millennium. This is a point-value because there is (we may fairly suppose) just one world where this happens. Once it becomes clear just what having point-values would be like, the anti-Humean may fairly complain that the conception of rational agency at work is too idealised. Even if he concedes that the fully rational agent has point-values, the sensible anti-Humean surely did not mean to restrict his claim to such Übermenschen.[19] And if the desire-as-belief thesis is only a requirement for less-than-fully-rational agents, that may be good enough.

Second, why think that constancy is permissible anyway? Surely the anti-Humean's discovery that (under certain assumptions) constancy is not rationally permissible is at best a mildly surprising fact, not a reductio of DAB.

Lewis disagrees. He holds, not just that constancy is rationally *permissible*, but that it is rationally *required*, on the basis of the following argument:

> If [*W*] were maximally specific merely in all "factual" respects relevant to its value, and if the Desire-as-Belief Thesis were true, then it would be no surprise if a change in belief changed our minds about how good it would be that [*W*], and thereby affected the value of [*W*]. But the subcase [i.e. *W*] was supposed to be maximally specific in *all* relevant respects—and that includes all relevant propositions about what would and would not be good. The subcase has a maximally specific hypothesis about what would be good built right into it. So in assigning it a value, we do not need to consult our opinions about what is good. We just follow the built-in hypothesis ... It is unintelligible how a shift in opinions about what is good could affect the value of any of the maximally specific [*W*]'s ... . (1988, p. 332)[20]

A point-proposition *W* is maximally specific in all respects, being true at exactly one possible world. (So, for vividness, we may think of *W* as a

[19] Of course, we could say the same about beliefs in point-propositions.

[20] The quotation in fact concerns a proposition labelled "AE_iF_h" that is "maximally specific in all respects relevant to its value" (1988, p. 332). A point proposition *W* is a special case of such. (Lewis is concerned with a more general claim because he is arguing that DAB leads to absurdity under the assumption that Jeffrey conditionalising is a rationally permissible rule of updating.)

maximally consistent set of sentences.) A fortiori, it is maximally specific in evaluative respects: for any sentence of the form "*A* is good", *W* will specify whether it is true or false (assuming, as the anti-Humean does, that such sentences *do* have truth values). So, once the rational agent grasps *W*, there is just no wiggle room to vary the value he assigns to it, any more than he may vary his opinion about whether *W* implies that there are blue swans. So Lewis argues.

We reply as follows. It is highly plausible that all evaluative facts supervene on purely descriptive facts. Assume this is so, and that all propositions are either evaluative or purely descriptive. Then any worlds w_1, w_2 that are descriptively exactly alike are also exactly alike in respect of what is good, in which case $w_1 = w_2$, and so $W_1 = W_2$. Suppose an agent desires a proposition *A* that is maximally specific in all *descriptive* respects. Then, given the supervenience thesis, if *A* is true at w_1 and true at w_2, then $w_1 = w_2$. In other words, *A* is a point-proposition. Therefore it is possible to value a point-proposition *W* by grasping only its specification of descriptive matters. A rational agent may have such values. So it is possible rationally to value a point-proposition *W* and be unaware or unsure about what is good at the world at which it is true. And therefore changing one's mind about the value one assigns to point-propositions is not irrational.

To this it may be objected that the ideally rational agent will be a priori certain of exactly what supervenes on what; in particular, given the descriptive specification of some *w*, he will have worked out, with a priori certainty, exactly what is good at *w*.

Perhaps this is right. But now the conception of rational agency is truly hyper-idealised: the rational agent not only has point-values, but a mind-boggling stock of a priori certainties. Very well, such a hyper-rational agent cannot obey the desire-as-belief thesis. That is certainly interesting, but the sensible anti-Humean will protest that he had a much less demanding sense of rationality in mind, one that might intelligibly serve as a norm for all-too-human agents. Let Hume be right about hyper-rational agents; that is at best only a partial victory.

Moreover, it is a partial victory *only if* an agent's value function can be said to measure his degree of desire. We will raise two reasons for doubt.

As we have noted, the Humean needs only the weak thesis that constancy is rationally permissible. But the only good argument in the offing for the weak thesis is Lewis's, and that purports to show the strong thesis, that constancy is rationally required. Therefore we may assume, as the Humean ought to, the strong thesis. Now the strong thesis has two important implications. First, that the rational agent's new value function, after he updates on *A*, is still defined on *A*. Second, that the rational agent's new value function, after he updates on the negation of a point-proposition *W*,

is still defined on W. Each of these two consequences of the strong thesis provides a reason for doubting that V measures the agent's degree of desire, as follows.

First, it is a folk psychological platitude that the desire for A is (or, at least, rationally ought to be) extinguished when the agent comes to believe A with certainty. (In fact, all we need is the assumption that losing this desire is rationally *permissible*; but let's stick with the stronger claim.) If you desire that Lara be at the party, and you come to be sure that she is at the party, then your desire ought to disappear (typically, of course, others will take its place). So when an agent updates on A, his new value function applied to A should be undefined. But of course it is not: assuming constancy, the agent's new value function applied to A will either remain equal to $V(A)$ (if V is updated, with the requirement that $C_A = (C_A)_A$) or at any rate (with V ratio-weighted and employing some other method of updating) will be perfectly well-defined.[21] Any rational agent, according to decision theory, always values every proposition he believes with certainty. But rational agents, according to folk psychology, do not desire such propositions.[22]

Second, it is a folk psychological platitude that it is rationally *permissible* to lose the desire for A when the agent comes to believe $\sim A$ with certainty.[23] But, assuming constancy, for all *point*-propositions W, when an agent updates on $\sim W$, his new value function is still defined on W. According to decision theory, there are certain propositions that an agent ought to value even when he believes their negations with certainty. But, according to folk psychology, an agent may reasonably lose his desire in any proposition of whose falsity he is certain.

Hence an agent's decision-theoretic values are dubious candidates to be identified with his desires. To the extent this is so, David Hume gets no help from decision theory.[24]

[21] Cf. Jeffrey (1983, pp. 62–3).

[22] An agent is *indifferent* to A just in case $V(A) = V(\sim A)$. It might be objected that the loss of the desire for A when the agent comes to believe A with certainty should not be represented as the agent's acquiring a value function that is *undefined* on A, but rather as the agent's becoming *indifferent* to A. This suggestion seems wrong: although there certainly is a psychological state of desiring A and $\sim A$ to the same degree, it does not appear to be the state an agent ends up in after strongly desiring A and then coming to believe it with certainty. For one thing, the suggestion does not account for the disposition of the agent to regain his strong desire for A, under the condition of losing his conviction in A. In any case, the difficulty is not evaded: it is usually not true that when an agent comes to believe A with certainty, $V(A) = V(\sim A)$.

[23] Perhaps such loss of desire is not rationally required. One may *wish* that some event—that one is sure occurred—had not happened, without thereby being irrational. And, depending on how exactly the anti-Humean thinks of desires, maybe wishes are kinds of desire.

ALEX BYRNE
Department of Linguistics and Philosophy
Massachusetts Institute of Technology
20D-213
Cambridge, MA 02139
USA
abyrne@mit.edu

ALAN HÁJEK
Division of the Humanities and Social Sciences
101-40
California Institute of Technology
Pasadena, CA 91125
USA
ahajek@hss.caltech.edu

REFERENCES

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

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

——1991: *Belief Revision.* PhD dissertation, Princeton University.

Gärdenfors, Peter 1982: "Imaging and Conditionalization". *Journal of Philosophy,* 79, pp. 747–60.

Gibbard, Allan, and Harper, William 1978: "Counterfactuals and Two Kinds of Expected Utility", in C. A. Hooker, J. J. Leach, and E. F. McClennen (eds.), *Foundations and Applications of Decision Theory,* vol. I, Dordrecht: D. Reidel.

Hájek, Alan 1994: "Triviality on the Cheap?", in Ellery Eells and Brian Skyrms (eds.), *Probabilities of Conditionals.* Cambridge: Cambridge University Press.

——1996: "The Fearless and Moderate Revision: Extending Lewis's Triviality Results", in Timothy Childers, Petr Kolář, and Vladimír Svoboda (eds.), *Logica 95*, Prague: The Institute of Philosophy of the Academy of Sciences of the Czech Republic.

Hájek, Alan, and Pettit, Philip1996: "In the Spirit of Desire-as-Belief", MS.

Hume, David 1739–40: *A Treatise of Human Nature.* L. A. Selby-Bigge (ed.), second edition, Oxford: Oxford University Press, 1978.

Jeffrey, Richard 1965: *The Logic of Decision.* second edition, Chicago: University of Chicago Press, 1983.

Lewis, David 1976: "Probabilities of Conditionals and Conditional Probabilities", in Lewis 1986. Originally published in 1976 in *Philosophical Review*, 85, pp. 297–315.

[24] Many thanks to John Collins, Ned Hall, David Lewis, Robert Stalnaker, and an anonymous referee for *Mind.*

——1979: "Counterfactual Dependence and Time's Arrow", in Lewis 1986. Originally published in 1979 in *Noûs*, 13, pp. 455–76.
——1981: "Causal Decision Theory", in Lewis 1986. Originally published in 1981 in *Australasian Journal of Philosophy*, 59, pp. 5–30.
——1986: *Philosophical Papers,* vol. II. Oxford: Oxford University Press.
——1988: "Desire as Belief". *Mind*, 97, pp. 323–32.
——1989: "Dispositional Theories of Value". *Proceedings of the Aristotelian Society, Supplementary Volume*, 63, pp. 113–37.
——1996: "Desire as Belief II". *Mind*, 105, pp. 303–13.
Price, Huw 1989: "Defending Desire-as-Belief". *Mind*, 98, pp. 119–27.
Stalnaker, Robert 1968: "A Theory of Conditionals", in Nicholas Rescher (ed.) *Studies in Logical Theory*. Oxford: Oxford University Press.
Teller, Paul 1973: "Conditionalization and Observation". *Synthese*, 26, pp. 218–58.