Counterfactual desire as belief

J. Robert G. Williams
j.r.g.williams@leeds.ac.uk

January 19, 2010

Bryne & Hájek (1997) argue that Lewis's (1988; 1996) objections to identifying desire with belief do not go through if our notion of desire is 'causalized' (characterized by causal, rather than evidential, decision theory). I argue that versions of the argument go through *on certain assumptions about the formulation of decision theory*. There is one version of causal decision theory where the original arguments cannot be formulated—the 'imaging' formulation that Joyce (1999) advocates. But I argue this formulation is independently objectionable. If we want to maintain the desire as belief thesis, there's no shortcut through causalization. [1]

Background

Lewis (1988, 1996) gave an impossibility proof for the identification of desires for p (or more specifically: having a certain degree of desire for p) with *belief* in a related proposition—which we might think of as expressed by 'it is good that p'. The proof might be turned to several ends, but Lewis intended it to refute one specific version of an 'anti-Humean' theory of motivation, on which certain doxastic states play the 'motivational' role that Humeans reserve for pro-attitudes.[2]

One way to think about decision theory has it characterizing 'belief-type roles' and 'desire-type roles' that jointly rationalize action. Within the theory, the former are represented by 'credence functions'; the latter are represented by 'expected value functions'. But the formalism itself does not fix what interpretation we give to these devices. For all we've so far said, sometimes the desire-role might be played by certain beliefs; or indeed sometimes the belief-role may be undertaken by certain desires.

The anti-Humean Lewis has in mind (in the simplest formulation) says that certain creatures that play the desire-type role are really *nothing but belief*. And so Lewis takes them to be saying that certain beliefs—beliefs about how good a prospect is—*play the decision theoretic desire-role*. What it *is* to desire that p (to a certain degree) is to believe that p is good, to that

[1]Thanks to all with whom I've discussed this material, including Elizabeth Barnes, Ross Cameron, Daniel Elstein, David Etlin, Graham Oddie, Mark Schroeder, and Jason Turner, The work in this paper was funded by a British Academy Research Development Award (BARDA: 53286).

[2]For an introduction to Humean and anti-Humean theories of motivation, see Smith (1994). Lewis's version is focused on decision-theoretic analogues of the 'all-or-nothing' formulations Smith talks of.

degree.[3] In particular, to change our degree of belief in that proposition is, *ipso facto*, to change our level of desire. We cannot tolerate the belief about goodness failing to play the desire-role in any coherent mental state.

Lewis's argument is sold, not least by Lewis himself, on its anti-Humean impact. Anti-Humeans may protest that the position Lewis attacks is a straw man—not something they were committed to. Some obvious initial concerns can be assuaged by making the model more complex, but how best to represent anti-Humean thought in decision theoretic terms will no doubt remain controversial. Lewis himself emphasizes that many anti-Humeans will feel themselves to have reason to resist Lewis's characterization.

However, seeing the debate in these confrontational terms at all—as an attempted 'refutation' of a particular position, with the associated rhetoric of 'straw men', 'mischaracterizations' and so forth seems to me unhelpful. I prefer to see the enterprise as follows. Anti-Humeans are committed to some kind of interesting relationship between desire and belief. Many accept that decision theory captures systematic relationships between these notions. So a natural way to study the anti-Humean thesis is to find a way to express it in terms of the 'desirabilities' and 'credences' of decision theory. Connecting these two areas of philosophy would be very interesting, since ideas and arguments could then flow back and forth across the bridge—not least, mathematical tools could be brought to bear to study the consequences of the anti-Humean thesis.

The impossibility result is then an obstacle in the way of this hopeful prospect of collaboration—the most natural starting point for putting an anti-Humean thesis into decision theoretic terms looks like it collapses. This would be bad news for everybody. Anti-Humeans lack a way of translating their views (even in a simplified toy model) to the decision theorists; and decision theorists would lack the capacity to study decision making under anti-Humean assumptions. If we could find a consistent reformulation (or some principled way of resisting the impossibility argument itself, while maintaining the equation) we could resurrect the bridge.[4]

It is not only anti-Humeans who should take interest. For the connection between motivation and belief about the good is one of the central features of metaethical debate, and some analogue of desire as belief is a prima facie commitment of many metaethical views. Take those expressivists who might wish to maintain that what we tendentiously describe as a belief that p is good or 'desirable', really is an expression of a desire that p. For roughly opposite reasons to the anti-Humeans, it seems they will want to assert an identity between beliefs and desires. One gloss that they might give is that there's *nothing more* to believing that p is good, than *valuing* p in a certain kind of way.But again, if we're to link beliefs and desires in this way, it looks like we cannot tolerate the desire failing to play the corresponding belief-role in any coherent mental state. To relate this to the 'beliefs' and 'desires' of decision theory, it

[3]There's an ambiguity in 'believing that p is good to a degree'—according to whether the degrees are degrees of belief or degrees of goodness. Lewis can accommodate both.

[4]One consistent reformulation is put forward by Broome (1991). One can maintain that desires match a certain expectation-value defined entirely in terms of one's belief state (including one's credences in the goodness of *complete outcomes*). If we only required value-judgements pertaining to *complete outcomes* to be motivational, then the Broome formulation gives us what we want. It's natural to want more—to consider the motivational role of *believing that giving this money to charity is good*, for example. But the availability of the Broome version should allay fears that the decks are being stacked against the anti-Humean from the start.

looks like we should investigate the desire as belief thesis.

Oddie suggests another way of looking at the identity. Cognitivists quite generally allow there to be genuine beliefs about what prospects are or are not good. In general, they needn't think there is a necessary connection between beliefs and desires—no belief need be intrinsically motivating. But as Oddie notes, cognitivists may very well want to hold out the ideal of a well-functioning individual—who not only holds true beliefs about the good, but who *harmonizes* their motivations to their value-judgements—i.e who desires something just to the extent she believes it good. Oddie suggests cognitivists quite generally should be be concerned iwth the alleged failure of the desire as belief thesis, since it articulates conditions under which the ideal of harmony is satisfied. Again, we view this constructively: what is the ideal of harmony between motivation and value-judgement? Can we capture it precisely using utilities and credences? The most natural suggestion, it seems, gets into trouble.

What then is this 'impossibility result'? Lewis's proof (in its simplest form) shows that *if we assume that coherent belief states are closed under certain natural 'conditionalizations'* the *systematic* identity between belief and desire cannot be maintained. Specifically, starting from a mental state in which one's degree of belief that p is good matches one's level of desire (expected value) in p, we can show that the belief that 'p is good' won't play the desirability/expected-value-of-p role in a mental state that arises from a single step of 'updating' from the first.[5] Importantly, this basic form of a 'desire as belief' thesis generalizes. Impossibility results can be extended, for example, to cases where we allow the 'goodness' of prospects to come in *degrees*, and demand that desires match the what one *expects* the goodness of the prospect to be (Lewis, 1988). The basic version is a harmless simplification.

There are two obvious places where morals can be drawn, if Lewis's formulation of desire and belief is indeed refutable. We could try to use it to learn something about metaethics—"better not formulate the link between motivation and value-judgement *that* way!". This, of course, was the sort of significance Lewis had in mind. The second is to use something about the utilities and credences of decision theory—"better not think of updating as working *that* way". I believe the best response is the second—-specifically that our view of updating should be made more sophisticated to allow a general *class* of cases of which desire as belief is one.[6] However, in the remainder of the paper I explore an alternative response of this second type, one suggested by Bryne & Hájek (1997) among others.[7] This accepts (at least for the sake of argument) Lewis's entire framework, but says that Lewis's argument misfires because he chooses the *wrong decision-theoretic representation of desirability*—in particular, that he should be working with the causalist 'expected utility' rather than the evidentialist 'expected value'.

[5]If there must be a necessary connection between (coherent) belief and desire, it's obvious what the relevance of this result is. For Oddie's cognitivist, it's not quite so straightforward—for there is nothing as yet that suggests the the identity between belief and desire cannot be achieved *statically, i.e. within a single mental state*. The worry would have to be that Lewis's result shows that harmony would have to be overly fragile—disappearing under the impact of the least information.

[6]We should not confuse the Bayesian thesis that updating one's credences is must always proceed by conditionalization, from the very strong thesis that updating by an arbitrary proposition A is always rationally permissible—it may be that there are some propositions where are not possible 'total information' one could receive. Lewis's argument as stated requires the latter. But a more interesting challenge (that I explore elsewhere) is to give a characterization of what updating should be, given the prima facie receipt of factual information p.

[7]They ascribe a similar view to John Collins, in his PhD thesis. Oddie (2001) suggests a similar moral. Weintraub (2007) notes the possibility of evading Lewis's result by moving to a different decision theoretic framework.

1 Lewis's argument

Lewis's argument is the following. First, we take 'off the peg' a characterization about how credence P, value u and expected value V interrelate:[8]

$$V(A) = \sum_w P_A(w)u(w)$$

Here $P_A(X)$ is the credence one has in X *supposing* A *to be actual*, a notion standardly characterized (at least when $P(A) > 0$) as the ratio $P(AX)/P(A)$.

Lewis thinks of V as describing the *desirability* of the proposition A. Hence when A° represents the content *that* A *is good*, the desire-as-belief thesis is characterized as:[9]

$$V(A) = P(A^\circ)$$

Let's be specific. u—the agents 'preferences among complete situations' is assumed to remain constant. But her beliefs and desires can change on receipt of new information. The above equation is supposed to hold of all coherent states, so in particular:

Now Lewis takes it that if V, P is a coherent combination, the result of 'receiving total information A' should also be coherent. On receiving total information A, our new credences $P'(X) = P(X|A)$. Since we hold the basic desirability of complete outcomes, described by u, fixed, we can read off the new 'desirabilities' V' by the equation given above. So by the desire as belief hypothesis, applied twice, we have:

$$V(A) = P(A^\circ)$$
$$V'(A) = P'(A^\circ)$$

Notice something rather special about this choice of update. We have $P'_A(X) = P_A(X|A) = P_A(X)$ (the last identity holds because $P_A(X) = P(X|A)$—so conditionalizing on A once again makes no difference). Thus we can see:

$$V'(A) = \sum_w P'_A(w)u(w) = \sum_w P_A(w)u(w) = V(A)$$

So updating by conditionalization gives you $P'(X) = P(X|A)$. Crucially, for *this particular update* we have $V(X) = V'(X)$. Putting this together with the desire-as-belief thesis, we get:

[8] This is Jeffrey-style 'evidential decision theory'. See (Jeffrey, 1967).

[9] This is just the most simple version that Lewis discusses. We could, for example, replace the all-or-nothing 'desirability' operator with degree-desirability operators.

$$P(A^\circ|A) = P'(A^\circ) = V'(A) = V(A) = P(A^\circ)$$

But P was an arbitrary coherent credence. Thus we can argue that for any coherent credence P, $P(A^\circ|A) = P(A^\circ)$. This is the thesis that Lewis calls 'independence' IND. And he then shows that it itself cannot be sustained through a class of credences closed under conditionalization—and by Lewis's assumptions that coherent credences are closed under conditionalization, this would show that, contrary to our assumptions, belief that A is good cannot play the corresponding 'desire' role in every rational state of mind.

To close out the proof, therefore, we need to show that IND can't hold under conditionalization. To do this, we think a little bit about what happens when we update by $A \vee A^\circ$. Except in trivial cases, it turns out that we have the following:[10]

$$P(A^\circ|A) = P_{A\vee A^\circ}(A^\circ|A)$$
$$P(A^\circ) < P_{A\vee A^\circ}(A^\circ)$$

We now cannot equate the two LHS and the two RHS, as IND would require. So contrary to our assumptions, IND isn't true for every coherent credence. QED.

I do not find the argument watertight. The natural place to resist is in the assumption that coherent credences are closed under conditionalization.[11] It is *not* obvious to me that we can get 'pure' information about factual matters like A, whilst leaving entirely unchanged our views about whether each proposition is all things considered good or bad. Even if we think that learning goes by conditionalization, we needn't say that every possible conditionalization is a coherent way of learning something.[12] But the topic for this paper is a different line of response, suggested by Bryne & Hájek (1997).

[10]This works provided P assigns positive probability to each of $A \wedge A^\circ$, $\neg A \wedge A^\circ$, $A \wedge \neg A^\circ$, $\neg A \wedge \neg A^\circ$. But if desire as belief can only be maintained in cases where one of these possibilities is ruled out completely, surely the game is already up.

[11]That is, we might regard it as a *constraint of rationality* that judgements about the good have the motivational role they do. The debate here is very close to that over the 'dynamic' satisfiability of 'Ramsey' conditionals—as Lewis himself notes (Lewis, 1988).

[12]Andrew McGonigal has suggested a second way of resisting at the same point. In circumstances when the motivational role is missing, we might say that the *sense* of 'good' differs—so we simply *do not express the same concept* in those states of minds where the desire as belief hypothesis fails. (It would probably be best to formulate this with a version of anti-Humeanism, or expressivism, which did not *eliminate* or *reduce* beliefs about goodness to desires or vice versa, but rather just maintained that a possession condition for the full-fledged *concept* of the good that judgements about that concept play the motivational role captured by the identity.

When the motivational constraints are not met, on this picture we do *not* token the same concept of 'good' that we do in the original cases. But we might use the very same word 'good' to express the thoughts we do have, and it might have the very same semantic value (i.e. both pick out the Lewisian proposition A°. This would be a form of *disjunctivism* about the concept of the good—'good' shorn of its motivational underpinnings is a related, but different, concept from the full-fledged version. In this version, conditionalization takes us to a *rationally faultless* mental state, but one that involves *different and less rich concepts*.

2 Two notions of desirability

It is intriguing that Lewis identifies desirability with expected *value* (calculated via evidential decision theory) rather than expected *utility* (calculated via causal decision theory). The characterization of the former is given above, whereas the latter is given by:

$$U(A) = \sum_w P^A(w)u(w)$$

Here $P^A(X)$ is the probability of X under the 'supposition as counterfactual' of A. To get a fix on this idea, consider same of the classic situations from the literature on conditionals. Under the supposition that Oswald *actually didn't* shoot Kennedy, someone else must have. But under the *counterfactual* supposition that Oswald *hadn't* shot Kennedy, Kennedy would probably have survived.

Characterizing causal decision theory (CDT) in terms of counterfactual supposition in the first instance is not the standard approach (though see Joyce (1999) for advocacy of a supposition treatment). Elstein & Williams (manuscript) defend the view that the suppositional formulation is the most basic form of decision theory, as an articulation of (instrumental) practical rationality. But we will not make this assumption here—we will instead consider various *candidate characterizations* of P^A.

Lewis's view is that CDT is a theory of *choiceworthyness* of various prospects, and EDT is a theory of *desirability* of those same prospects. The two can come apart, he says: in a 'Newcombe' puzzle we *desire* to one-box; but the choiceworthy option is to two-box. The desirability of one-boxing has some phenomenological plausiblity. Suppose one woke up the next day, having forgotten what choice you made. What choice would you *hope* you made (caring only about money?) Many would say that one-boxing is the thing to hope for here—it is the better *news* to receive.

Bryne & Hájek (1997) object to the divorcing of desirability and choiceworthiness. And there does seem something odd with the idea that one has an all-things-considered desire that A occur, has it totally in one's power to bring it about that A happen; but yet it is rational to bring about $\neg A$. It seems to me that the CDT theorist faces a genuine tension here: either deny a plausible desirability-choiceworthyness connection, or deny what seems like data about what it is rational to hope for.[13]

Bryne & Hájek (1997) suggest friends of CDT as a theory of practical rationality (choiceworthyness) should also endorse it as an account of desirability. This undermines Lewis's arguments against desire as belief: for what it is to 'play the desire role' is now to (globally) match expected utility U, rather than expected value V. And when we try to run the original argument in the new setting, we note that it *no longer* works as formulated. For a crucial piece of the argument was that when we updated our credences by conditionalization on A, the expected utility calculation ensured that $V(A) = V'(A)$. And this in turn rested on the

[13] Another proposal is that we diagnose ambiguity: in a hoping/wishing sense we desire 1-boxing; but in some 'genuine desire' sense relevant to agency we desire 2-boxing. See Etlin.

fact that the 'weights' that enter into the characterization of $V(A)$ are themselves probabilities conditional on A, and hence unchanged when we update by that very proposition. This piece of the argument now lapses. Hence, Bryne & Hájek (1997) suggest, desire-as-belief thesis is tenable on a causalized notion of desire extracted from CDT.

3 Resurrecting the impossibility argument

As sketched above, Lewis's argument against desire as belief is blocked when we have a *causalized* notion of desire in place, because conditionalization isn't guaranteed to preserve $U(A)$, as it was to preserve $V(A)$. A first step to reintroducing the impossibility result is to see what sort of change to the credences would reintroduce this step.

The key thing we need, in choosing our pair P, $\bar{P}$, is that $\bar{P}^A(X) = P^A(X)$; just as previously we choose P, P' such that $P'_A(X) = P_A(X)$. Here is a natural thought: let $\bar{P}$ simply be the probability function P^A (just as we let P' be P_A). This choice will satisfy our principal so long as the following plausible constraint holds of counterfactual supposition—reiterations of the same supposition have no effect (all the various articulations below satisfy this). Given $\bar{P}^A = (P^A)^A = P^A$, we can then run the argument:

$$\bar{U}(A) = \sum_w \bar{P}^A(w)u(w) = \sum_w P^A(w)u(w) = U(A)$$

So we have $\bar{P}(X) = P^A(X)$ but also, crucially, $U(X) = \bar{U}(X)$. Putting this together with the desire-as-belief theses:

$$P(A^\circ) = U(A)$$
$$\bar{P}(A^\circ) = \bar{U}(A)$$

we get:

$$P^A(A^\circ) = \bar{P}(A^\circ) = \bar{U}(A) = U(A) = P(A^\circ)$$

Hence we can conclude that for *any* coherent starting credence P, $P^A(A^\circ) = P(A^\circ)$. This is an analogue of the problematic 'independence' result. But having reached this point, we are stuck—we need to appeal to further resources to continue the parallel to the original proof.

As stated at present, there's an obvious reason why we can prove impossibility in the case of EDT, but not in the case of CDT. In the case of EDT, we have independent formal traction on P_A—via the ratio formula. At the moment, P^A hasn't been given a formal characterization (except for the one, plausible requirement we needed to assume above). This doesn't mean it's ill-understood—counterfactual supposition is a familiar kind of state, I contend—but we need to say something more about its formal behaviour before we can go further.

But what if we help ourselves to the richer characterizations implicit in popular formulations of decision theory? The easiest to conceptualize is the following *conditional principal principle*:[14]

$$P^A(X) = \sum_x x \cdot P(Ch(X|A) = x)$$

This has considerable appeal—as much, I think, as the unconditional version of the Principal Principle; and as much as the identification of $P_A(X)$ with $P(AX)/P(A)$.

Assuming that fully rational agents obey the conditional principal principle, we can consider the consequences of the above for the special case where agents are fully informed of the chances. Their credences will then match the (known) chances/conditional chances, and so:

$$P^A(A^\circ) = Ch(A^\circ|A)$$
$$P(A^\circ) = Ch(A^\circ)$$

From our earlier argument, we have $P^A(A^\circ) = P(A^\circ)$. Hence, by the above, we derive:

$$Ch(A^\circ|A) = Ch(A^\circ)$$

But now notice that this is a version of the *original* independence identity—just with a different interpretation of probability: chances rather than credences.

Now re-run the above argument for the chance function $Ch_{A\vee A^\circ}$—assuming this *is* a possible chance function. We get another instance of chance-independence:

$$Ch_{A\vee A^\circ}(A^\circ|A) = Ch_{A\vee A^\circ}(A^\circ)$$

Just as in Lewis's original case, we can see from the behaviour of probabilities under conditionalization that (except in trivial cases)

$$Ch_{A\vee A^\circ}(A^\circ|A) = Ch(A^\circ|A)$$

$$Ch_{A\vee A^\circ}(A^\circ) > Ch(A^\circ)$$

But these are inconsistent with the pair of 'chance-independence' equations. Contradiction, and QED.

We have worked here with a characterization of P^A by chance. Another widespread characterization is extractable from the K-partition formulation of CDT. With K a privileged partition, consisting of exhaustive and exclusive 'causal dependency hypothesis', we identify:

[14] We will have to adjust this (perhaps by conditionalization on total evidence) in order to deal with cases of 'inadmissible information'. But for now, let's just stick to this as a restricted identity, holding at least for agents whose information is admissible.

$$P^A(X) = \sum_{k \in K} P(k)P(X|Ak)$$

Now take the special cases where the agent is certain which dependency hypothesis is actual. Then the above simplifies to:

$$P^A(X) = P(X|A)$$

So for these very special agents, P^A and P_A coincide. We can thus show that $P(A^\circ|A) = P(A^\circ)$. A knowledgable agent might then come to be certain that $A \vee A^\circ$, and hence $P_{A \vee A^\circ}(A^\circ|A) = P_{A \vee A^\circ}(A^\circ)$—as is now familiar, we can derive inconsistency from this pair. So both of these characterizations of (CDT and thus) P^A are reducible to absurdity via Lewis's impossibility results, suitably adapted.

4 A consistent desire as belief thesis

Not every formulation of CDT is susceptible to an adaption of the desire as belief argument. Recall that $P^A(A^\circ) = P(A^\circ)$. If we read A° as $A \mathrel{\square\!\!\rightarrow} G$ for some proposition G, then this becomes:

$$P^A(A \mathrel{\square\!\!\rightarrow} G) = P(A \mathrel{\square\!\!\rightarrow} G)$$

Is there a reading of P^A on which the above, so construed, is tenable?

Let us assume that $\mathrel{\square\!\!\rightarrow}$ is a Stalnaker conditional. Then it is well known that the *imaging* probability of P under A, defined by $I(P,A)(\bullet) := P(A \mathrel{\square\!\!\rightarrow} \bullet)$, is a probability function. And by defining $P^A(X) = I(P,A)(X)$, we have in general:

$$P^A(B) = P(A \mathrel{\square\!\!\rightarrow} B)$$

So imaging probabilities (relative to a counterfactual Stalnaker conditional) are a *possible* reading of counterfactual supposition. Let's consider what this reading says about the probability of $A \mathrel{\square\!\!\rightarrow} G$ under A. We have:

$$P^A(A \mathrel{\square\!\!\rightarrow} G) = P(A \mathrel{\square\!\!\rightarrow} (A \mathrel{\square\!\!\rightarrow} G))$$

But the proposition on the RHS is equivalent in the Stalnaker logic to $A \mathrel{\square\!\!\rightarrow} G$. So we have:

$$P^A(A \mathrel{\square\!\!\rightarrow} G) = P(A \mathrel{\square\!\!\rightarrow} G)$$

But we have now *derived* the problematic equation above. Since G was arbitrary, it seems that so long as we read A° as $A \boxdot\!\!\rightarrow G$, we won't run into trouble.

5 Evaluation

I have argued that on two prominent formulations of CDT, desire cannot be belief—or at least, the thesis is just as problematic as in Lewis's original setting. In one formulation of CDT—one based on a Stalnaker counterfactual, and the induced imaging probabilities—it can be sustained.

This might seem odd, for one often gets the impression that the various formulations of CDT are equivalent or intertranslatable. Thus a famous quote from Lewis: "We causal decision theorists share one common idea, and differ mainly on matters of emphasis and formulation" (Lewis, 1981, p.5). Skyrms and Harper say: "It can be argued that the various forms of causal decision theory are equivalent— that an adequate version of any one of them will be interdefinable with adequate versions of the others". Joyce (1999, p.173) maintains: "there is no great difference between any of the approaches we have discussed. . . the causal decision theorist can adopt an attitude of benign indifference". But there is a strong reading of such claims that is highly contentious (and I do not think would be endorsed by the above authors).[15] For example in order for the chance formulation and the counterfactual formulation to *give the same recommendations in decision problems*, we would need their characterizations of P^A to coincide, and thus:

$$P(A \boxdot\!\!\rightarrow B) = \sum_x x \cdot (P(Ch(B|A) = x))$$

However, I argue elsewhere (Williams, manuscript) this identity between the probability of a counterfactual and corresponding expected conditional chance, is just as problematic as the notorious putative identity between probability of indicative conditionals and the corresponding conditional probability.[16]

If the formulations are *inequivalent* in the sense of giving different recommendations for action, then we have a choice to make—and causal decision theorists who start in *different* places can be seen as plumping for one recommendation over the other. For example, Joyce

[15] One immediate wrinkle is that Lewis and Joyce, at least, will ditch the *counterfactual* formulations for the closely related *generalized imaging* formulations—but as I'll explain below imaging formulations diverge from the others we have considered.

[16] Lewis and Joyce would in any case not endorse the above—their favoured 'counterfactual' formulation of decision theory is in terms of 'generalized imaging' rather counterfactuals. This is presumably one reason for the hedges in the above statements. But the deeper point carries through. As Joyce proves (Joyce, 1999, ch.5) generalized imaging is an upper bound for the probability of the corresponding counterfactual, and so to get the imaging and chance formulations to line up, they would need a version of the above where the inequality $\leq$ replaces the $=$ (a Ramsey *bound* on the counterfactual, rather than a Ramsey *identity*. And this is no better than the equality—I argue in (Williams, manuscript) that it is susceptible to the very same (Lewisian) impossibility proof.

Briggs (manuscript) argues, on similar assumptions, that we can reintroduce analogues of many of the impossibility proofs described in Hájek & Hall (1994)

(1999) prominently supports the imaging formulation—*identifying* it with the notion of counterfactual supposition. If I am right then someone who took counterfactual supposition to be normed by conditional chance will give *different* recommendations for action.[17] This gives us leverage for evaluation.

Let us evaluate. To begin with, the following should be common ground: the imaging formulation of CDT gives hostages to fortune, in that it *defines* P^A in terms of a certain counterfactual. We must wait nervously for best theory of counterfactual conditionals to tell us what confidence to have in various central cases, and hope that the results are sensible by decision theoretic lights.[18]

Extant theories give us more reason to be nervous. For example, Lewis (1979) and Williams (2008) argue for views on which, it turns out, we can be credence 1 in a counterfactual $A \mathrel{\square\!\!\rightarrow} B$, even though the conditional chance of B on A is (known to be) lower.[19] This means that the corresponding *imaging* probability of B on A is higher than the conditional chance.[20] For example, Lewis and Williams would say that the following is flatly true, when the antecedent is false:

> If we flipped a fair coin a billion times, it wouldn't land all heads

Of course, they would agree that the conditional chance of not-all-heads on billion-flips is less than this (albeit not by much). Hawthorne (2005) points to more dramatic cases of this—cases where we can be certain of the counterfactual, even though the corresponding conditional chance is only $1/3$. Now, you might (with Hawthorne) regard these as *objections* to the theories of counterfactuals under discussion. But if the story about counterfactuals work approximately ok in the majority of cases, and only yields very surprising results in some recherche examples, then its not clear that the objection is at all decisive (Williams (2008) argues they are not). Furthermore, one appealing to such considerations is in an awkward position. For if they insist that in all cases, expected conditional chance should bound rational credence in a counterfactual, they are imposing a 'Counterfactual Ramsey Bound' that leads to essentially the same problems as the more famous probabilistic constraints on indicatives. And at this point, it seems the objector should really be telling us what they think a theory of counterfactuals looks like that meets such constraints. On the other hand, if they evade this by admitting that in some cases such a Bound is violated, why should it be problematic that it is violated in *these particular cases*—shouldn't we just steel ourselves against such complaints?

[17] Notice that if we *denied* the equation of P_A with conditional probabilities, and instead identified them with the imaging probability of a putative Stalnakerian indicative conditional, we could run the above tenability argument for desires as belief within EDT as well. See Bryne & Hájek (1997) for discussion.

Presumably the reason that *this* didn't seem to many a decent response to the desire-as-belief thesis in the first place is that an 'imaging' version of EDT just didn't seem right—the conditional version seemed the right way to go. And the very same thing can, I think, be said about the imaging formulation of CDT.

[18] Alternatively, we might say that 'being decision-theoretically sensible' is a *constraint* on theorists of the counterfactual. But as will become clear, I think that a decision-theoretically sensible conditional will have to be one for which imaging matches conditional chance in exactly the way that leads to impossibility results.

[19] At least in circumstances where we are certain the antecedent is false

[20] Joyce proves that the imaging probability is at least bounded below by the probability of the counterfactual.

But bringing such results into decision theory by insisting on an imaging-formulation is surely disastrous. Suppose hell results if you flip all-heads in some long sequence of coin flips, if you play the devil's game. If you play the devil's game and don't get all-heads, you get two smarties. If you don't play the game, you just get one smarty. The conditional chance of getting two smarties on playing the game is very high (and known to be so) but not 1. And weighting by chance, the disvalue of the hell outcome swamps the value of two smarties. Chance-based causal decision theory tells you not to run the risk. Imaging based causal decision theory gives no weight at all to the 'hell possibility', and so tells you to play the game, despite your conscious awareness of the immense disvalue lying in wait.

In Hawthorne's example, the case is even more dramatic. There are cases where you think playing the devil's game is *more objectively likely than not* to bring you disvalue, but since you can 'counterfactually exclude' this more probable outcome, imaging based causal decision theory tells you to play it.

This sort of thing just seems the *wrong* results to me—causal decision theory formulated in the imaging way is a bad guide to action. One could plausibly maintain, focusing solely on counterfactuals, we had little choice but to bite some bullets, in light of impossibility results. But this is not the case for friends of CDT—for the K-partition and chance formulations stand in line to give you the *intuitively correct* recommendations.

Now the *specific* examples I just gave trade on *specific* theories of counterfactuals. But unless imaging coincides (ideally) with expected chance, I think it's *inevitable* we're going to have objections of this kind.

Exactly because it comes apart from the conditional chances, I think Joyce's favoured imaging formulation is no good as a formulation of CDT. But this now removes the remaining hope of evading Lewisian arguments against desire as belief: for on the plausible formulations of causal decision theory, as we have already seen, they go through.

6 Where we end up

Pace Bryne & Hájek (1997), causal decision theory brings no new comfort for the defender of desire as belief. The Lewis arguments can be reformulated against such a position, and the only version of causal decision theory that promises an evasion of the results is independently objectionable. Of course, maybe a friend of desire as belief will anyway like the causal decision theory framework—to avoid having to distinguish between desirability and choiceworthyness—but this would be on *independent* grounds, not because it gives them an easier time defending the equation.

References

Briggs, Rachael. manuscript. 'Two interpretations of the Ramsey Test'.

Broome, John. 1991. *Weighing Goods*. Oxford: Blackwell.

Bryne, A., & Hájek, A. 1997. 'David Hume, David Lewis, and decision theory'. *Mind*, **106**(423), 423–428.

Elstein, Daniel Y., & Williams, J. Robert G. manuscript. 'Suppositions and decisions'.

Hájek, Alan, & Hall, Ned. 1994. 'The hypothesis of the conditional construel of conditional probability'. *In:* Eels, E., & Skyrms, B. (eds), *Probability and Conditionals*. Cambridge: CUP.

Hawthorne, John. 2005. 'Chance and counterfactuals'. *Philosophical and Phenomenological Research*, 396–405.

Jeffrey, Richard. 1967. *Formal logic: its scope and limits*. Third edn. New York: McGraw-Hill.

Joyce, James M. 1999. *The foundations of causal decision theory*. Cambridge, UK: Cambridge University Press.

Lewis, David K. 1979. 'Counterfactual Dependence and Time's Arrow'. *Noûs*, **13**, 455–76. Reprinted with postscript in Lewis, *Philosophical Papers II* (Oxford University Press, 1986) 32–51. Also reprinted in Jackson (ed) *Conditionals* (Oxford University Press, 1991) 46-76.

Lewis, David K. 1981. 'Causal Decision Theory'. *Australasian Journal of Philosophy*, **59**, 5–30. Reprinted with postscript in Lewis, *Philosophical Papers II* (Oxford University Press, 1986) 305-36.

Lewis, David K. 1988. 'Desire as belief'. *Mind*, **97**, 323–332. Reprinted in Lewis, *Papers on Ethics and Social Philosophy* (Cambridge University Press, 2000) 42–54.

Lewis, David K. 1996. 'Desire as belief II'. *Mind*, **105**, 303–313. Reprinted in Lewis, *Papers on Ethics and Social Philosophy* (Cambridge University Press, 2000) 55–67.

Oddie, Graham. 2001. 'Hume, the BAD paradox, and value realism'. *Philo*.

Smith, Michael. 1994. *The Moral Problem*. Oxford: Blackwell.

Weintraub, Ruth. 2007. 'Desire as belief, Lewis notwithstanding'. *Analysis*.

Williams, J. Robert G. 2008. 'Chances, counterfactuals and similarity'. *Philosophy and Phenomenological Research*, **77**(2), 385–420.

Williams, J. Robert G. manuscript. 'A Lewis-impossibility result for counterfactuals'.