Objective Value Is Always Newcombizable

Arif Ahmed and Jack Spencer

Words: 5885

Words, including footnotes and appendices: 9060

This essay argues that evidential decision theory (EDT) is incompatible with options having objective values.

After some scene-setting (§§1-2), we consider three arguments for our thesis: the argument from Newcomb's Problem (§§3-4), the argument from Expectationism (§§5-6), and the argument from Newcombizability (§7). The first two arguments fail for instructive reasons. But the third succeeds. EDT is inconsistent with options having objective values because objective value is always Newcombizable.

§1/ Objective and Subjective

We begin by supposing:

> **Two Oughts**. There is both an objective *ought* and a subjective *ought*.

One can deny Two Oughts — Mackie (1977), for instance, rejects the objective *ought*, and Moore (1903) and Thomson (2008) reject the subjective *ought*.[1] But cases like the following convince many philosophers to accept it:[2]

> *Miners*. Ten miners are trapped in shaft A or shaft B, but the agent does not know which. Flood waters threaten to flood the shafts. The agent has enough sandbags to block one shaft, but not both. If the agent blocks one shaft, all the water will go into the other, killing any miners inside. If the agent blocks neither shaft, both shafts will fill halfway with water, and just one miner, the lowest in the shaft, will be killed. (In fact, the miners are trapped in shaft A.)

With regard to Miners, one might ask: what ought the agent to do? And is the agent's uncertainty relevant to what she ought to do? If Two Oughts is true, these questions lack univocal answers.

In the *objective* sense, the agent's uncertainty is irrelevant. However confident she is that the miners are in shaft B, she objectively ought to block shaft A. In the *subjective* sense, the uncertainty is paramount. If the agent is 5% confident that the miners are in shaft A and 95% confident that they are in shaft B, then she subjectively ought to block shaft B. If, as we suppose henceforward, the agent is 50% confident that the miners are in shaft A and 50%

[1] Also see Kolodny and MacFarlane 2010, which argues for only one *ought* that is neither purely objective nor purely subjective.

[2] Parfit (unpublished); *cf.* Jackson 1991 and Regan 1980. For a different motivation of Two Oughts, see Oddie and Menzies 1992.

confident that they are in shaft B, she subjectively ought to block neither shaft — despite knowing for certain that blocking neither shaft is not what she objectively ought to do.

Given the distinction between objective and subjective *ought*, we can say what it would be for options to have objective and subjective *values*.

Saying that options have *objective* values is equivalent to saying that some consequentializable theory of the objective *ought* is true. In other words, it's to say that there is a numerically representable property of options such that, whenever an agent faces a decision, the options that are objectively permissible for her to choose are exactly the options that maximize the property.

Some philosophers conjecture that every (plausible) theory of the objective *ought* is consequentializable.[3] Although we will not assume that this conjecture is true, if it is, our arguments establish an even more surprising metaethical thesis — that EDT is incompatible with every (plausible) theory of the objective *ought*.

Saying that options have *subjective* value is equivalent to saying that some consequentializable theory of the subjective *ought* is true. Like objective values, subjective values are numerically representable properties of options. But unlike objective values, subjective values are had only relative to a credence function. So to claim that options have subjective values is to claim that there is some numerically representable property of options relative to a credence function such that, whenever an agent faces a decision, the options that are subjectively permissible for the agent to choose are exactly the options that maximize the property relative to the agent's credence function.

[3] See *e.g.* Dreier 1993; 2011, Louise 2004, Portmore 2007.

To understand why EDT is inconsistent with options having objective values, it will be helpful to contrast it with its chief rival, causal decision theory (CDT).

§2/ EDT and CDT

EDT and CDT are both consequentializable theories of the subjective *ought*. But they disagree over what subjective value *is*.

According to EDT, subjective value is *evidential expected value*. To characterize evidential expected value, we need some familiar formalism.

Let $W = \{w_1, w_2, \dots, w_m\}$ be the set of *possible worlds*. We assume for simplicity that W is finite. And to ease the formalism, we ignore the distinction between a possible world and its singleton.

Let $A = \{a_1, a_2, \dots, a_n\}$ be the set of the agent's *options*, each of which is a proposition (i.e. a set of possible worlds) that the agent can make true by deciding.[4] We assume that options are always pair-wise exclusive and jointly exhaustive.

Let C be the agent's *credence function*, a probability function mapping each proposition to a real number in the unit interval and thereby representing the agent's confidence that the proposition is true. Let $C(p)$ be the agent's credence in p. And in the usual way, let $C(p|q) = C(pq)/C(q)$ be the agent's *conditional* credence in p given q.

Finally, let u be the *world-valuation function* mapping each possible world to a real number and thereby representing the value of that world. Value comes in many flavors, and

[4] Following Jeffrey 1983: 83-4.

for our purposes any flavor will do. If we are interested in morality, we may take $u(w)$ to represent how morally good w is. If we are interested in prudential value, we may take $u(w)$ to represent how prudentially good (i.e. desirable) w is; and so on. Our claim that EDT is incompatible with *options* having objective values holds for any flavor of value that might belong to individual worlds.

With this formalism in place we can characterize evidential expected value. The evidential expected value of a, written $V(a)$, is:

Evidential Expected Value: $V(a) = \sum_W C(w|a)u(w)$.[5]

According to CDT, subjective value is *causal expected value*. The simplest way to characterize causal expected value, which, for convenience, we adopt here, involves a *highly discerning causal similarity metric*.[6] That means: a measure of distance between possible worlds (a *similarity metric*) such that: (i) it is *highly discerning*: for any option a and any world w_i, some a-world w_j is, by the lights of this measure, *uniquely* closest to w_i[7]; (ii) it is

[5] For simplicity, this characterization of evidential expected value involves credences in individual worlds, thus departing from the technical assumption in standard axiomatizations of EDT that preference is defined over an atomless field (Jeffrey 1983: 146). This is harmless, since nothing that we say depends on the representation theorem for EDT that requires atomlessness for its proof; in any case, we could do without atoms as long as value is not 'gunky' (as Hájek (2013: 438) puts it).

[6] For discussion of other characterizations, see e.g. Ahmed 2014, ch.2, Joyce 1999 ch. 5, Lewis 1981.

[7] No natural similarity metric is quite this discerning: for discussion see Lewis 1973 and Joyce 1999 sect. 5.4. A more realistic proposal is that, for any option a and world w_i, there is some set of a-worlds, $\{w_j, w_k, \ldots, w_m\}$ that are, by the lights of the causal similarity metric, more similar to w_i than are any other a-worlds. Nothing except ease of exposition is going to turn on this.

causal: it reckons the closest a-world to w_i to match w_i on all particular matters of fact to which a is causally irrelevant.

Given this highly discerning causal similarity metric, we can give truth-conditions for nonbacktracking counterfactual conditionals. If $\langle a \Rightarrow w_j \rangle$ is the nonbacktracking counterfactual — that world w_j would have been actual had option a been chosen — then $\langle a \Rightarrow w_j \rangle$ is true at some possible world w_i if and only if w_j is the a-world that is, by the lights of the highly discerning causal similarity metric, uniquely most similar to w_i.

The causal expected value of option a, written $U(a)$, is a function of the agent's credences in these nonbacktracking counterfactuals:

Causal Expected Value: $U(a) = \sum_W C(\langle a \Rightarrow w \rangle)u(w)$.

§3/ The Argument from Newcomb's Problem

The first purported argument for our thesis is *the argument from Newcomb's Problem*. EDT and CDT often agree on which option maximizes subjective value, but not always. One famous crux is:

> *Newcomb's Problem*. There is a transparent box and an opaque box. The agent has two options: she can take only the opaque box (she can 'one-box'), or take both boxes ('two-box'). The transparent box contains $1,000. The opaque box contains either $1,000,000 or $0, depending on a prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would two-box, then the opaque box contains

> $0. If the predictor predicted that the agent would one-box, then the opaque box contains $1,000,000. The agent knows all of this.

The agent reasonably takes her choice to be strong evidence about whether the opaque box contains $1,000,000 or $0. Equating units of value and dollars, the V-score of one-boxing is therefore slightly under 1,000,000, and that of two-boxing slightly over 1,000.[8]

But two-boxing uniquely maximizes U. Interpreting the following counterfactuals in terms of a causal similarity metric, the agent's credence that she would have received $1,000 had she two-boxed exactly equals her credence that she would have received $0 had she one-boxed, and her credence that she would have received $1,001,000 had she two-boxed exactly equals her credence that she would have received $1,000,000 had she one-boxed. The U-score of two-boxing is therefore exactly 1,000 greater than that of one-boxing, no matter what the U-score of one-boxing is.

The argument from Newcomb's Problem refers to Newcomb's Problem and has two premises:

> **Newcomb Knowledge**: If options have objective values then: an agent facing Newcomb's Problem knows for certain that two-boxing uniquely maximizes objective value.

[8] E.g. if the agent expects the predictor to be accurate on 99% of trials whichever option is chosen, then the V-score of one-boxing is $(0)(0.01) + (0.99)(1{,}000{,}000) = 990{,}000$; and that of two-boxing is $(1{,}000)(0.99) + (0.01)(1{,}001{,}000) = 11{,}000$.

> **Bridge**: If options have objective values then: if an agent's options are $a_1, a_2, \ldots, a_n$, and the agent knows for certain that option a_i uniquely maximizes objective value, then option a_i uniquely maximizes subjective value.

Newcomb Knowledge is motivated by the claim that objective value is *causal value*. The causal value of an option is the value of the world that would have been actual had the agent chosen the option. That is: if $\langle a \Rightarrow w_j \rangle$ is true at world w_i, then the causal value of a at w_i is $u(w_j)$. Causal value is a familiar conception of objective value. It features, for example, in Moore's defense of consequentialism, where he writes:

> In order to show that any action [maximizes objective value], it is necessary to know both what are the other conditions, which will, conjointly with it, determine its effects; to know exactly what the effects of these conditions will be; and to know all the events which will be in any way affected by our actions throughout an infinite future. We must have all this causal knowledge…. And not only this: we must also possess all this knowledge with regard to the effects of every possible alternative; and must then be able to see by comparison that the total value due to the existence of the action in question will be greater than that which would be produced by any of these alternatives. (1903: 149)

It is uncontroversial that an agent facing Newcomb's Problem knows for certain that two-boxing uniquely maximizes causal value. So, if objective value is causal value, Newcomb Knowledge is true.

Bridge is motivated by metaethical considerations. The subjective *ought* arises from the agent's subjective uncertainty about the objective *ought*. When an agent is uncertain what she objectively ought to do, there can be a discrepancy between what the agent subjectively and objectively ought to do. Indeed, as Miners illustrates, when an agent is uncertain about she objectively ought to do, an agent can know for certain that they differ. But if an agent knows for certain what she objectively ought to do, then the objective and subjective *ought* must coincide.[9] Two Oughts thus entails the following conditional: *if an agent knows for certain that she objectively ought to choose option a_i, she subjectively ought to choose option a_i*.[10] And given that options have both objective and subjective values, this conditional is equivalent to the consequent of Bridge.[11]

[9] Zimmerman defends the stronger claim that '[a]n agent [subjectively] ought to perform an act if and only if he believes that it is [objectively] the best option that he has' (2008: 5). We think Miners refutes the left-to-right direction of this claim; but the right-to-left direction is plausible and entails Bridge.

[10] One apparent challenge to Bridge involves upward monotonicity. The agent in Miners objectively ought to block shaft A. If the objective *ought* is upward monotonic, then the agent in Miners will (if rational) know for certain that she objectively ought to block some shaft or other. But it is not the case that the agent subjectively ought to block some shaft or other. In response one might reasonably deny upward monotonicity; but in any case, the example does not threaten Bridge as stated, because Bridge only covers an agent's *options*, and since options are pair-wise exclusive, *blocking some shaft or other* cannot be an option if *blocking shaft A* is an option.

[11] Bridge must be distinguished from nearby principles.

One nearby principle concerns permissibility. It says: if an agent knows for certain that an option is objectively *permissible*, then the option is subjectively *permissible*. This principle is plausible, but less plausible than Bridge. Opaque sweetening cases, discussed in Hare 2010, put pressure on this permissibility principle, but put no pressure on Bridge.

It might appear that the argument from Newcomb's Problem establishes our thesis — that Nozick, way back in 1969, showed that EDT is inconsistent with options having objective values. But the argument from Newcomb's Problem can be resisted.

§4/ Many Conceptions of Objective Value

Bridge is undeniable, but Newcomb Knowledge is not. An agent facing Newcomb's Problem knows for certain that two-boxing uniquely maximizes causal value. But why should EDT'ists grant that objective value is causal value? After all, there are *many* conceptions of objective

Another nearby principle concerns inevitable knowledge. It says: if an agent knows for certain that she would be objectively required to choose a particular option if she knew that *p* was true and also knows for certain that she would be objectively required to choose that same option if she knew that *p* was false, then the agent is subjectively required to choose the option. This inevitable knowledge principle is clearly false. It might be the case that an agent, who is uncertain about whether *p*, objectively ought to pay a small sum of money to come to learn the truth-value of *p*, even though the agent knows for certain that she would be objectively required not to pay the small sum if she knew that *p* was true or knew that *p* was false.

The difference between Bridge and the principle concerning inevitable knowledge is relevant to the discussion in Hare 2016. Hare makes two claims: (1) that killing one to save five is subjectively permissible when it is unknown which of the six is being sacrificed to save the other five, and (2) that killing one to save five is subjectively impermissible if the identity of the one being sacrificed to save the five was known. It is not obvious to us that both claims are true. But even if they are, they pose no threat to Bridge. We claim that if (1) is true, then it is also objectively permissible to kill one to save five when the identity of the one being sacrificed to save the five is unknown. And we claim that if (2) is true, then it is also objectively impermissible to kill one to save five when the identity of the one being sacrificed to save the five is known. Hare's example might be a counterexample to the inevitable knowledge principle, but it is not a counterexample to Bridge.

value (where by ‘conception of objective value’ we mean a proposal about what objective value is).

A natural conception of objective value has two components. The first is a similarity metric. The input to a similarity metric is some option a and some world w_i, and the output is some set $\{w_j, w_k, \dots, w_m\}$ of a-worlds that are, on that metric, the most similar a-worlds to w_i. The second component is a method of averaging. The objective value of option a at world w_i is some weighted average of $u(w_j), u(w_k), \dots, u(w_m)$, the values of the a-worlds that are, by the lights of the similarity metric, most similar to w_i. If a similarity metric always outputs a set containing only one possible world, we can disregard the method of averaging, for the objective value of an option a at world w_i is then just the value of the a-world that is, by the lights of the similarity metric, uniquely most similar to w_i. Causal value is *one* member of this family of natural conceptions of objective value, but there are many others. And many are uniquely maximized by *one-boxing* in Newcomb’s Problem.[12]

For instance, consider what Horgan says, in his defense of one-boxing:

> I shall assume that there is indeed a standard resolution of the vagueness of the similarity relation among worlds, and that Lewis’s account of it is essentially correct. Returning to Newcomb’s problem, it is clear that [the] premises…of the one-box

[12] The full class of objective value concepts extends well beyond this natural family. In its broadest sense, an objective value concept is *any* function taking each pair of a possible world and a proposition to a real number. Although we find it more natural and intuitive to focus on the restricted family that the main text describes, our result extends to any objective value concept within this broader class. We claim that EDT rules out identifying objective value with *any* function within that broader class.

> argument cannot…be true under the standard resolution. For the being made his prediction about my choice, and has either put the $1 million in the [opaque box] or not, well before I choose. Thus, his *actual-world* prediction and the *actual-world* state of [the opaque box] remain intact in the closest world in which I take both boxes, and also in the closest world in which I take [the opaque box] only.
>
> [… The] intuitive plausibility of the one-box argument rests upon a nonstandard resolution, one that seems quite appropriate in this context. It differs from the standard resolution to the extent that it gives top priority to maintaining the being's *predictive correctness* in the nearest possible world where I take both boxes, and also in the nearest world where I take [only the opaque box]. Under this *backtracking resolution* … the closest world in which I take both boxes is one in which the being correctly predicted this and put nothing in [the opaque box], and the closest world in which I take only [the opaque box] is one in which he correctly predicted *this* and put $1 million in [it]. (1981: 336)

Although Horgan does not give details, we can spell out a similarity metric on possible worlds that satisfies his desiderata. One way to do it would be to graft Horgan's 'top priority' onto something like Lewis's (1979) lexicographic criteria for measuring the relative similarity of worlds, insisting that what counts most for closeness of a possible world *w* in this context is whether the predictor's accuracy at *w* with regard to the agent's choice matches the predictor's actual accuracy on that question.[13] We then use the Horgan metric to define a conception of objective value, which we might call:

> **Horgan value**: if w_j is the a-world that is most similar to w_i by the Horgan metric, then the Horgan value of a at w_i is $u(w_i)$.

[13] More formally, define a measure of the relative similarity of worlds w and w' to a fixed world w_i by appeal to five partial orders on worlds:

1) $w >_1 w'$ iff: the Newcomb predictor's correctness on this occasion at w matches the predictor's accuracy at w_i, but the predictor's correctness at w' does not.
2) $w >_2 w'$ iff: there are big, widespread and diverse violations of the laws of w_i at w' and not at w.
3) $w >_3 w'$ iff: the spatio-temporal region throughout which perfect match over particular facts with w_i prevails is larger at w than at w'.
4) $w >_4 w'$ iff: there are small, localized, simple violations of the laws of w_i at w' and not at w.
5) $w >_5 w'$ iff: w achieves approximate similarity to w_i over matters of particular fact and w' does not.

Say that w_j is *Horgan-closer* to w_i than is w_k if and only if: either (i) $\{n | w_k >_n w_j\} = \emptyset$ and $\{n | w_j >_n w_k\} \neq \emptyset$; or (ii) $\min\{n | w_k >_n w_j\} > \min\{n | w_j >_n w_k\}$.

Criteria 2)-5) are from Lewis (1979: 47-8), minus Lewis's hedging over 5). Taken jointly as an analysis of the 'standard resolution' of closeness in natural language, 2)-5) are implausible (see e.g. McDermott 1999 for criticism). Of course we are not making any claims about natural language counterfactuals but rather using Lewis's criteria to *construct* a kind of objective value that one-boxing maximizes.

Often, the Horgan value of an option is equal to its causal value. For instance, equating lives and units of value, in Miners, the causal value of blocking shaft A (B) is 10 (0), and likewise the Horgan value of blocking shaft A (B) is 10 (0).

But in Newcomb's Problem*,* the two diverge. The causal metric holds fixed the contents of the opaque box but not the predictor's correctness. Whatever money is in the opaque box at the two-boxing-world that is causally closest to the actual world is also in the opaque box at the one-boxing-world that is causally closest, so the causal value of two-boxing exceeds that of one-boxing by 1,000. The Horgan similarity metric holds fixed the predictor's correctness but not the contents of the opaque box. So, if the predictor is actually correct, the two-boxing-world that is Horgan-closest is one where the opaque box contains $0, and the one-boxing-world that is Horgan-closest is one where the opaque box contains $1,000,000. If the predictor is actually incorrect, the situation is reversed — the two-boxing-world that is Horgan-closest is one where the opaque box contains $1,000,000, and the one-boxing-world that is Horgan-closest is one where the opaque box contains $0. In either case, the Horgan value of some option diverges from its causal value.

If objective value is Horgan value, Newcomb Knowledge is false. An agent facing Newcomb's Problem does *not* know for certain that two-boxing uniquely maximizes Horgan value. After all, she is confident that the predictor is accurate on this occasion, so she is confident that one-boxing uniquely maximizes Horgan value. Indeed, if objective value is Horgan value and the predictor is known to be sufficiently reliable, the agent might know for certain that *one-boxing* uniquely maximizes objective value. And Horgan value is not unique in this respect. Uncountably many conceptions of objective value can be known to be uniquely maximized by one-boxing in Newcomb's Problem.[14]

The argument from Newcomb's Problem is therefore too quick. The argument establishes *something*. Since Bridge is true, the argument establishes that EDT is inconsistent with any conception of objective value that validates Newcomb Knowledge. But most conceptions of objective value invalidate Newcomb Knowledge. Without some independent argument that the true conception of objective value validates Newcomb Knowledge, the argument from Newcomb's Problem fails to establish our thesis.

§5/ The Argument from Expectationism

The second purported argument for our thesis is *the argument from Expectationism*.

Expectationism is a thesis about how objective value and subjective value relate. It says that subjective value is expected objective value. Let $\langle O(a) = v \rangle$ be the proposition that (on some arbitrarily chosen scale) the objective value of an option a is $v \in \mathbb{R}$. Then, more formally, we have:

[14] Thus consider the view on which the objective value of an option a is given by the conditional chances at some time before the decision. Suppose the prediction is at t_1, and consider the conditional chances tagged to some earlier t_0. We can formulate a similarity metric that makes the closest a-worlds to w_i be those in the set of worlds $\{w_j, w_k, \dots, w_m\}$ that have positive chance at time t_0 at w_i, conditional on a. To calculate the objective value, average the values of the outputted worlds by their respective chances conditional on a. Thus, according to the proposal, the objective value of a at world w_i equals $\sum_W Ch_{w_i,t_0}(w|a)u(w)$. If the predictor has a 99% chance of correctness back at t_0, then the objective value of one-boxing is $(0)(0.01) + (1{,}000{,}000)(0.99) = 990{,}000$, and the objective value of two-boxing is $(1{,}000)(0.99) + (0.01)(1{,}001{,}000) = 11{,}000$.

Expectationism: For any credence function C, the subjective value of a relative to C is $\sum_v vC(\langle O(a) = v\rangle)$.

Expectationism is widely accepted. Often it's just assumed,[15] but there are some explicit defenses of it. For example, one might defend Expectationism by claiming that the subjective value of an option should be the agent's best estimate of its objective value, and then arguing that the agent's best estimate of a quantity is their expectation of it.[16]

In the dispute between one-boxers and two-boxers, Expectationism is neutral. If we combine Expectationism with the claim that objective value is causal value, we get CDT, since the causal expected value of an option is the agent's expectation of the causal value of the option.[17] But we can combine Expectationism with conceptions of objective value that invalidate Newcomb Knowledge. For example, if we combine Expectationism with the claim that objective value is Horgan value, then we get *Horgan decision theory* (HDT) — the view that the subjective value of an option is the agent's expectation of the Horgan value of the option. Whereas two-boxing uniquely maximizes expected causal value (i.e. causal expected value), one-boxing uniquely maximizes expected Horgan value.

[15] See e.g. Parfit 1984: 25.

[16] For a defense of Expectationism along these lines, see *e.g.* Oddie and Menzies 1994 and Pettigrew 2015.

[17] *Proof:* Let $O_w(a)$ be the causal value of option a at world w. The expectation of the causal value of a relative to credence function C is $\sum_W C(w)O_w(a)$. Let W_i be the worlds at which $\langle a \Rightarrow w_i\rangle$ is true. Then $\sum_W C(w)O_w(a) = \sum_{W_1} C(w)u(w_1) + \cdots + \sum_{W_n} C(w)u(w_n) = C(\langle a \Rightarrow w_1\rangle)u(w_1) + \cdots + C(\langle a \Rightarrow w_n\rangle)u(w_n) = \sum_W C(\langle a \Rightarrow w\rangle)u(w) = U(a)$.

But Expectationism is hostile to EDT, as we now argue.

Every remotely plausible conception of objective value must allow an agent to regard a as evidence about what the objective value of a is. Take the simplest case, in which an agent is uncertain whether the objective value of option a is v_1 or v_2. Every remotely plausible conception of objective value must validate:

Relevance: It is possible that an agent's credences be such that, for some $v_1 \neq v_2$:

i. $C(\langle O(a) = v_1 \rangle \vee \langle O(a) = v_2 \rangle) = 1$,

ii. $C(\langle O(a) = v_1 \rangle) = x < 1$, and

iii. $C(\langle O(a) = v_1 \rangle | a) = y \neq x$.

Relevance must hold because the fact that an agent regards a as evidence as to whether p cannot *preclude* the objective value of a from depending on whether p.

To see this more concretely, consider a variation on Miners. On this variation, the agent remembers the miners' telling her which shaft they would be working in, but cannot consciously recall which. As before, she is 50% confident that the miners are in shaft A and 50% confident that they are in shaft B. But she (reasonably) thinks that there is a nonzero chance that her unconscious memory will influence her choice if she chooses to block one of the shafts. Therefore, her confidence that the miners are in shaft A, conditional on her blocking shaft A, is (say) 52%, up from 50%.

Since clauses (i) and (ii) of Relevance clearly hold in this case, anyone who denies Relevance would have to hold that the objective value of blocking shaft A does not depend on where the miners are. They would have to hold that the objective value of blocking shaft

A at a world at which the miners are in shaft A is equal to the objective value of blocking shaft A at a world at which the miners are in shaft B. But that's absurd. On any remotely plausible conception of objective value, the objective value of blocking shaft A depends on where the miners actually are, even if the agent regards blocking shaft A as evidence about where the miners are. Thus, Relevance must hold.[18]

Now we can prove:

Result #1. Relevance, Expectationism, and EDT are jointly inconsistent.

The (very simple) proof is in Appendix A. Informally, the idea is that Expectationism makes the subjective value of an option depend *only* on the agent's unconditional credences in its having this or that objective value, whereas EDT makes it turn on its expected objective value *conditional on its performance*. Relevance therefore implies that EDT and Expectationism can disagree about the subjective value of an option. For instance, in the variant on Miners,

[18] In saying this, we are effectively setting aside 'indexical' value concepts of the sort that Hájek and Pettit (2004) discuss in connection with Lewis's (1988, 1996) arguments against 'Desire as Belief' (DAB). Indexical values depend not only on the state of the mind-independent world but also on the beliefs of the agent herself. We agree with Hájek and Pettit that indexical value concepts evade Lewis's arguments. They also violate Relevance. (For instance, if the indexical value of an option a is just $V(a)$, then, so long as conditionalization can change the evidential expected value of the tautology (see Bradley and Stefánsson 2016: 699-702), Relevance as applied to indexical value fails.) But we deny that indexical value is *objective.* The notion of objective value that interests us is such that, if options have objective values, the objective value of (say) blocking shaft A in Miners depends on where the miners *actually* are, regardless of what the agent thinks. (For more on DAB see n. 23.)

Expectationism implies that the subjective value of blocking shaft A is 5, whereas EDT reckons it at 5.2.

The argument from Expectationism exploits Result #1. It says that EDT is inconsistent with options having objective value because, if options have objective values, Relevance and Expectationism are both true.

The argument from Newcomb's Problem was too narrow. It showed that EDT is inconsistent with any conception of objective value that validates Newcomb Knowledge, but was silent about conceptions of objective that do not validate Newcomb Knowledge. The argument from Expectationism, by contrast, purports to establish that EDT is inconsistent with every (remotely plausible) conception of objective value, even those, like Horgan value, that invalidate Newcomb Knowledge. It might seem, then, that the argument from Expectationism establishes our thesis — that a proper understanding of the mathematical relationship between subjective value and objective value reveals that EDT is inconsistent with options having objective values. But the argument from Expectationism can be resisted.

§6/ Resisting Expectationism

Every remotely plausible conception of objective value validates Relevance, so the argument from Expectationism establishes that EDT and Expectationism are inconsistent. Evidential expected value is not expected objective value. CDT'ists and HDT'ists agree that subjective value is expected objective value and disagree about what objective value is, but EDT'ists do not share in this agreement. *No* conception of objective value stands to EDT as causal value stands to CDT.

But Expectationism can be questioned. If options have both objective and subjective values, then there must be *some* well-behaved, intimate relationship between the subjective value of an option relative to an agent's credence function and the agent's hypotheses about its objective value. But this relationship needn't be expectation.

A minimal necessary condition for the relationship between objective and subjective value being well-behaved and intimate is the truth of the following principle:

> **Certain Reflection**: For any credence function C, if $C(\langle O(a) = v\rangle) = 1$, then the subjective value of option a relative to C equals v.

But Certain Reflection is strictly weaker than Expectationism, as we can see by considering some views that verify Certain Reflection while falsifying Expectationism.

First consider *Maximin* — the view that the subjective value of a relative to an agent's credence function is the least v such that the agent assigns nonzero credence to $\langle O(a) = v\rangle$. Maximin satisfies Certain Reflection, but it falsifies Expectationism.

Next, consider *Risk-adjusted Expectationism* — the view that the subjective value of a is not the straight expectation of objective value but rather a weighted sum of its possible objective values that attaches more importance to some of these possibilities than to others depending on their rank and not only on their probability: for instance, it weights worse possibilities more heavily than better ones, other things being equal.[19] On most such weightings, the view that subjective value is risk-adjusted expectation of objective value

[19] Buchak 2013 ch. 2.

falsifies Expectationism, but again it satisfies Certain Reflection, because if you are certain of what the objective value of an option is, then there *are* no alternative hypotheses about this to which you can give more or less weight depending on their rank.[20]

A third alternative, which is amenable to EDT, takes subjective value to be *conditional expected objective value*. We call this:

> **Conditional Expectationism**: For any credence function C, the subjective value of a relative to C is $\sum_v vC(\langle O(a) = v\rangle|a)$.[21]

Conditional Expectationism entails Certain Reflection. And given Certain Reflection, EDT entails Conditional Expectationism.[22]

[20] We can realize this idea formally by means of a *distortion* i.e. a non-decreasing function $r: [0,1] \rightarrow [0,1]$ such that $r(0) = 0$, $r(1) = 1$. For any option a, arrange its epistemically possible objective values in increasing order $x_1, x_2, \dots, x_n$. The corresponding version of *Risk-adjusted Expectationism* (RE) says that relative to a credence function C the subjective value of a is $\sigma(a) = x_1 + \sum_{i=1}^{n-1}(x_{i+1} - x_i)r\big(C(\langle O(a) \geq x_{i+1}\rangle)\big)$. If r is the identity function i.e. $r(x) = x$ then subjective value coincides with expected objective value; but if r is convex to the x-axis, e.g. if $r(x) = x^2$, then subjective value weights worse possibilities more heavily than expectationism demands. But trivially, it satisfies Certain Reflection.

[21] Oddie 1994: 460. Also see Broome's (1991) discussion of Conditional Expectationism (which he somewhat misleadingly calls Desire-as-Expectation).

[22] Proof: Let w_i be any world at which a and $\langle O(a) = v\rangle$ are both true, and let C concentrate all of its credence on w_i. Then, by Certain Reflection, the subjective value of a relative to C is v. And, by EDT, the subjective value of a relative to C is $\sum_W C(w|a)u(w) = u(w_i)$. So $u(w_i) = v$. This means that EDT entails Conditional Expectationism, since, $\sum_{w \in W} C(w|a)u(w) = \sum_{w \in \langle O(a) = v_1\rangle \cap a} C(w|a)u(w) + \cdots = \sum_{w \in \langle O(a) = v_1\rangle} C(w|a)v_1 + \cdots = \sum_v vC(\langle O(a) = v\rangle|a)$.

There is something intuitive about Expectationism, which says that the subjective value of an option should be the agent's estimate of its objective value, and someone who accepts Conditional Expectationism must reject Expectationism, since, given Relevance, the two claims cannot both be true. But there is also something intuitive about Conditional Expectationism, which says that the subjective value of an option should be the agent's estimate of its objective value *in worlds where it is realized*. The theoretical cost of rejecting Expectationism in favor of Conditional Expectationism thus seems to us low. And there is no argument from Conditional Expectationism. Whereas Relevance, Expectationism, and EDT are jointly inconsistent, Relevance, Conditional Expectationism, and EDT are consistent.

The assumption that EDT'ists must accept Expectationism is unjustified. (Even among opponents of EDT, Expectationism is not common ground.) So, without some independent argument that EDT'ists must accept Expectationism, the argument from Expectationism fails to establish our thesis.[23]

[23] We should briefly relate the present discussion to Lewis's (1986, 1988) argument against the anti-Humean 'Desire as Belief' (DAB) thesis, to which the argument from Expectationism bears an obvious resemblance. The basic idea behind Lewis's argument is that if evidential expected value V measures the agent's desire for the truth of a proposition, then DAB says that $V(A) = C(\dot{A})$, where $\dot{A}$ is the proposition that it is good that A. (Lewis's proof involves an ungraded notion of goodness but could easily be extended to cover a graded notion, as the argument from Expectationism does.) Given Lewis's assumption that $V(A|A) = V(A)$ (which has been questioned: see Bradley and Stefánsson 2016: 699-702) it follows from DAB that $C(\dot{A}|A) = C(\dot{A})$; but this is inconsistent with the analogue of Relevance that Lewis implicitly assumes (1996: 309). One way for the anti-Humean to resist Lewis's argument (Price 1989: 122) would be to reformulate the anti-Humean thesis as 'Desire as *Conditional* Belief' (DACB), which says that $V(A) = C(\dot{A}|A)$. This thesis gives no traction to Lewis's argument, for just the same reason that Conditional Expectationism gives none to the argument from Expectationism. In response, Lewis argues (1996: 310-11) that DACB is a version of *desire by*

§7/ The Argument from Newcombizability

The third purported argument for our thesis is *the argument from Newcombizability.*

It starts from a weakening of Bridge. As we said, Bridge seems undeniable. If options have objective values then: if an agent's options are $a_1, a_2, \dots, a_n$, and the agent knows for certain that option a_i uniquely maximizes objective value, then a_i uniquely maximizes subjective value. But the subjective values of options relative to a credence function supervene on the agent's credences and the world valuations. We therefore can weaken Bridge, appealing to any of the agent's certainties, and not just those that constitute knowledge. The result is:

> **Dominance**. If options have objective values then: if an agent's options are $a_1, a_2, \dots, a_n$, and the agent is certain that option a_i uniquely maximizes objective value, then a_i uniquely maximizes subjective value.

necessity: it is committed to the existence of a proposition G which the agent desires to be true *whatever* her credences. We are not clear why this point has any force against DACB. Maybe the idea is that the anti-Humean thesis is a descriptive psychological thesis about a person's desires, and that it is simply false as a matter of fact that there is anything that everyone desires (cf. Lewis 1996: 304-5). However this may be, we are clear enough that no analogous point threatens Conditional Expectationism. After all, the latter is a *normative* thesis, because of the connection between subjective value and what one subjectively *ought* to do. Even if there is nothing that everyone does value, there might be something that everyone *should* value. In short, we think: (a) that the same objection arises against both Lewis's argument and the argument from Expectationism; and (b) that even if the former survives it, the latter does not.

Dominance is like Certain Reflection. Certain Reflection ensures that subjective value conforms to objective value *numerically* — it says that, if an agent is certain that the objective value of an option equals v, then, relative to the agent's credence function, the subjective value of the option also equals v. Dominance ensures that subjective value conforms to objective value *ordinally* — it says that, if an agent is certain that some option uniquely maximizes objective value, then, relative to the agent's credence function, the option also uniquely maximizes subjective value.

Dominance is entailed by many views about how objective and subjective values relate, including Minimax, Risk-adjusted Expectationism, CDT, HDT, and any form of Expectationism.[24] But Dominance is hostile to EDT. There is a generalization of Relevance —

[24] Proof that Expectationism implies dominance: let $O_w(a)$ be the objective value of option a at world w. If $C(\langle O(a) > O(b)\rangle) = 1$ and Expectationism is true, then, relative to C, the difference between the subjective value of option a and the subjective value of option b equals $\sum_W C(w)(O_w(a) - O_w(b))$. Since the agent is certain that the objective value of a exceeds that of b, at any world w to which C assigns nonzero credence, $O_w(a) - O_w(b)$ is positive. Hence, $\sum_W C(w)(O_w(a) - O_w(b))$ is positive, which entails that, relative to C, the subjective value of option a exceeds the subjective value of b. Proof that maximin implies dominance: let $\underline{a} = \min\{x | C(\langle O(a) = x\rangle) > 0\}$. If $\underline{a} \leq \underline{b}$ then $C\big(\langle O(a) \leq \underline{b}\rangle\big) > 0$, so $C(\langle O(a) \leq O(b)\rangle > 0$. Contrapositively, $C(\langle O(a) > O(b)\rangle) = 1$ implies $\underline{a} > \underline{b}$, so the maximin subjective value of a exceeds that of b. For the proof that Risk-adjusted Expectationism implies dominance, see Buchak 2013: 245-6. Note that this proof assumes that the distortion function r is strictly increasing. But even if we assume only that r is non-decreasing we can still prove the following weakened dominance principle: if an agent's options are $a_1, a_2, \dots, a_n$, and the agent is certain that option a_i uniquely maximizes objective value, then no *other* option uniquely maximizes subjective value. The argument from Newcombizability still goes through on this weakening of Dominance.

a principle that is strictly stronger than Relevance, but no less obviously true — that contradicts the conjunction of EDT and Dominance.

We will build up to the relevant principle in two stages. To start, suppose that the agent is certain that the objective value of option a_1 exceeds the objective value of option a_2 by some definite positive margin z, but is uncertain whether the objective values of a_1 and a_2 equal v_1 and $v_2 = v_1 - z$, respectively, or instead equal v_3 and $v_4 = v_3 - z$, respectively, where $v_1 - v_3 > z$. Any remotely plausible conception of objective value must validate:

Baseline Relevance. It is possible for the agent's credences to be such that:

i. $C(\langle O(a_1) = v_1 \rangle \vee \langle O(a_1) = v_3 \rangle) = 1$,

ii. $C(\langle O(a_1) = v_1 \rangle | a_1) = x < 1$,

iii. $C(\langle O(a_1) = O(a_2) + z \rangle) = 1$, and

iv. $C(\langle O(a_1) = v_1 \rangle | a_2) = y \neq x$.

The rationale behind Baseline Relevance is the same as that behind Relevance. Baseline Relevance holds because any remotely plausible conception of objective value must allow an agent to regard a_1 as evidence about what the objective value of a_1 is, even if the agent is certain that the objective value of a_1 exceeds the objective value of some other option a_2 by some margin z.

If Baseline Relevance holds of every remotely plausible conception of objective value, then so too does a variant in which clause (iv) is stronger than a bare inequality. Using the same notation, the principle is:

Newcombizability. It is possible for the agent's credences to satisfy:

i. $C(\langle O(a_1) = v_1 \rangle \vee \langle O(a_1) = v_3 \rangle) = 1$,

ii. $C(\langle O(a_1) = v_1 \rangle | a_1) = x < 1$,

iii. $C(\langle O(a_1) = O(a_2) + z \rangle) = 1$, and

iv. $C(\langle O(a_1) = v_1 \rangle | a_2) = y > x + \frac{z}{v_1 - v_3}$.

As far as we can see, *almost any* conception of objective value is Newcombizable.[25] Indeed, there seems to be a general recipe for Newcombizing. Let O be any conception of objective value. If there are options a_1 and a_2, then there are propositions $S_1 =_{\text{def.}} \langle O(a_1) = v_1 \wedge O(a_2) = v_1 - z \rangle$ and $S_2 =_{\text{def.}} \langle O(a_1) = v_3 \wedge O(a_2) = v_3 - z \rangle$. We can therefore construct a decision problem in which the payoffs are as follows:

	$\boldsymbol{S_1}$	$\boldsymbol{S_2}$
$\boldsymbol{a_1}$	v_1	v_3
$\boldsymbol{a_2}$	$v_2 = v_1 - z$	$v_4 = v_3 - z$

Table 1

[25] The reason for the qualification 'almost' is just this. (i) We are assuming that the objective value concept implies that objective value takes on the appropriate values for some options a_1 and a_2 in some possible worlds. Thus, for example, we disregard any conception of objective value on which every option has the same objective value at every world. (ii) We are assuming that any given distribution of objective values of options across possible worlds is consistent with any distribution of credences across those worlds. This does not seem contentious to us, since the *objective* value of options are at least sometimes totally insensitive to changes in the agent's credences. See the corresponding discussion of Relevance at n. 18.

We construct a credence function C on the atoms $\{a_i S_j\}_{i,j=1,2}$ as follows. Let $x = \frac{1}{2}\left(1 - \frac{z}{v_1 - v_3}\right)$. Let $y = \frac{1}{4}\left(3 + \frac{z}{v_1 - v_3}\right)$. Choose an arbitrary k, $0 < k < 1$. Let $C(a_1 s_1) = xk$, $C(a_1 s_2) = (1 - x)k$, $C(a_2 s_1) = y(1 - k)$ and $C(a_2 s_2) = (1 - y)(1 - k)$. Then it is easy to check that C satisfies all of clauses (i)-(iv) in the Newcombizability condition. If one needs a back-story, imagine that the mechanism that typically causes one to choose a_1 also and independently tends to promote a state S_2 in which both options possess less of whichever kind of objective value is at issue.[26]

[26] As a concrete illustration of this general result, note that Horgan value is Newcombizable. Suppose that you have a choice between taking box 1 and box 2, both boxes being opaque. Taking box 1 releases 2 units of welfare; taking box 2 does nothing. But a Newcombian predictor has written down what he expects your choice to be; and if you manage to outwit the predictor, you get a bonus of 8 units of welfare. Moreover, you know that the predictor is somewhat better at predicting people who choose box 1 (success rate of 0.625) than at predicting people who choose box 2 (success rate of 0.1875). This need not be because you are antecedently and robustly confident that the prediction is that you will choose box 1; it could be that activation of the part of your brain that inclines you to choose box 2 interferes with the predictor's brain-scanning device. (For a similar example see the 'Semi-Frustrater' problem in Spencer and Wells forthcoming.)

This case satisfies the four clauses of Newcombizability. (i) you are certain that the Horgan value of (a_1) taking box 1 is either $v_1 = 10$ (if the predictor is actually inaccurate) or $v_3 = 2$ (if the predictor is actually accurate). To spell out why: on the Horgan resolution of counterfactuals, the closest a_1-world to actuality, call it w_1, is one where the predictor is accurate if and only if he is actually accurate. So if the predictor is actually inaccurate then $u(w_1) = 10$; otherwise $u(w_1) = 2$. So certainly $O(a_1) = 10 \vee O(a_1) = 2$. (ii) Your confidence that the Horgan value of taking box 1 is 10, *given* that you take box 1, is $x = 0.375$. (iii) You are certain that the Horgan value of taking box 1 exceeds the Horgan value of (a_2) taking box 2 by $z = 2$ units, because if the predictor is actually accurate then the former is 2 and the latter is zero, and if the predictor is inaccurate then the former is 10 and the latter is 8. (iv) Your confidence that the Horgan value of taking box 1 is 10, given that

If every remotely plausible conception of objective value is Newcombizable, then Dominance and EDT cannot both be true, because, as we prove in Appendix B:

Result #2. Newcombizability, Dominance, and EDT are jointly inconsistent.

The intuitive idea behind this is as follows. Whatever objective value is, we can construct a case where an option a_1 has more of it than another option a_2 at every possible world; but the dominated option, a_2, is very good evidence that both options have *high* objective value. Dominance therefore demands that the subjective value of a_1 exceeds that of a_2; but EDT implies the opposite.

The argument from Newcombizability follows from Result #2. It says that EDT is inconsistent with options having objective values because, if options have objective values, Newcombizability and Dominance are true.

The argument from Expectationism can be resisted by retreating from Expectationism to Conditional Expectationism, but no analogous 'conditionalizing' maneuver helps against the argument from Newcombizability. To see this, consider a few proposals.

The obvious first proposal is to reject Dominance in favor of:

you take box 2, is $y = 0.8125 > x + \frac{z}{v_1 - v_3} = 0.625$. So the case represents a Newcombization of Horgan value.

> **Conditional Dominance 1**: If options have objective values then: if an agent's options are $a_1, a_2, \dots, a_n$, and the agent is certain that option a_i uniquely maximizes objective value given that a_i is realized, then option a_i uniquely maximizes subjective value.[27]

But this gets us no further, because Conditional Dominance 1 entails Dominance. If an agent is certain that a_i uniquely maximizes objective value, then the agent is also certain that a_i uniquely maximizes objective value given that a_i is realized. So, Newcombizability, Conditional Dominance 1, and EDT are jointly inconsistent.

A second proposal might be to reject Dominance in favor of:

> **Conditional Dominance 2**: If options have objective values then: if an agent's options are $a_1, a_2, \dots, a_n$, and the agent is certain that: the objective value of option a_i conditional on its realization exceeds the objective value of any other option a_k conditional on *its* realization, then option a_i uniquely maximizes subjective value.[28]

Conditional Dominance 2 is not inconsistent with the conjunction of EDT and Newcombizability, but that's because it says nothing meaningful at all. The objective value of an option is not had relative to a credence function, so there is no such thing as the objective value of an option conditional on its realization (or conditional on anything).

[27] Formally, we can write the consequent of Conditional Dominance 1: $C(\bigwedge_{k\neq i}\langle O(a_i) > O(a_k)\rangle \,|a_i) = 1 \rightarrow \bigwedge_{k\neq i} \sigma(a_i) > \sigma(a_k)$.

[28] Formally, we might try writing the consequent of Conditional Dominance 2 as follows: $C(\bigwedge_{k\neq i}\langle O(a_i|a_i) > O(a_k|a_k)\rangle) = 1 \rightarrow \bigwedge_{k\neq i} \sigma(a_i) > \sigma(a_k)$.

One could try to *give* ‘conditional objective value’ an objective meaning, perhaps by explicating it as a counterfactual conditional relative to some similarity metric. The objective value of an option conditional on its realization then would be the objective value that the option would have *were* it to be realized. This would give us:

> **Conditional Dominance 3**: If options have objective values then: if an agent’s options are $a_1, a_2, \dots, a_n$, and the agent is certain that: the objective value of option a_i were it to be realized exceeds the objective value of any other option a_k were *it* to be realized, then option a_i uniquely maximizes subjective value.[29]

But this just rearranges the deck-chairs. If O is a conception of objective value, then *the O-value of an option were it to be realized* is just *another* conception of objective value, and Conditional Dominance 3 is just Dominance asserted about it. This alternative conception of objective value is, like any remotely plausible conception, Newcombizable, so no progress is made. Newcombizability, Conditional Dominance 3, and EDT are jointly inconsistent.

There is no way around Dominance, just as there is no way around Certain Reflection. To reject either is to reject the whole idea of objective values — objective properties of options to which subjective values conform. Anyone who rejects Certain Reflection or Dominance rejects the idea that differences between objective and subjective values are always mere artifacts of the agent’s uncertainty about objective values.

[29] Formally, we might write the consequent of Conditional Dominance 3 as: $C\big(\forall n \bigwedge_{k \neq i}\colon (a_i \Rightarrow \langle O(a_i) = n \rangle) \rightarrow (a_k \Rightarrow \langle O(a_k) < n \rangle)\big) = 1 \rightarrow \bigwedge_{k \neq i} \sigma(a_i) > \sigma(a_k)$, where $\Rightarrow$ is the selected counterfactual operator.

What the argument from Newcombizability reveals is the real, *metaethical* lesson of Newcomb's Problem. In the fifty years since Nozick introduced the problem, there has been much ado about causation. Over and again, opponents of EDT make the same causal observations: the agent facing Newcomb's Problem has no control over the contents of the opaque box; the amount of money contained in the opaque box at the causally closest one-boxing world is likewise contained in the opaque box in the causally closest two-boxing world; the agent is in a position to know for certain that the causal value of two-boxing exceeds the causal value of one-boxing, and so on. These observations suggest that the lesson of Newcomb's Problem has something essentially to do with causation.

But it doesn't. Causal language is used because it is presupposed that objective value is causal value, but the metaethical lesson of Newcomb's Problem concerns the relationship between objective value and subjective value, on *any* plausible conception of objective value. Take *any* remotely plausible conception of objective value, be it causal or wholly noncausal. *That* conception of objective value will be Newcombizable. There will be a case where EDT recommends an option that the agent knows for certain to be objectively worse on that conception than the only alternative. The EDT'ist's claim that subjective value is evidential expected value thus will be inconsistent with the claim that *that* conception of objective value is true.

Maximin, Risk-adjusted Expectationism, and the various form of Expectationism, including CDT and HDT, are consistent with options having objective values. Indeed, arguably these theories are defensible *only if* options have objective values. But EDT is metaethically very different. EDT is *not* consistent with options having objective values.

§8/ Conclusion

Past this point, the authors of this paper part ways. We agree that EDT is inconsistent with options having objective value, but we don't agree about what to make of that fact. One us of has antecedent commitments to options having objective values, so is inclined to take the inconsistency to amount to a metaethical refutation of EDT. The other has antecedent commitments to EDT, so is inclined to take the inconsistency to amount to a metaethical refutation of the claim that options have objective values. Obviously, we cannot settle here whether to reject EDT or the claim that options have objective value. But one of them has got to go.

Appendix A

Here we prove Result #1: that Relevance, Expectationism, and EDT are jointly inconsistent. Start from a possible case that witnesses the truth of Relevance. Then, according to the agent's credences, $C(\langle O(a) = v_1\rangle \vee \langle O(a) = v_2\rangle) = 1$, $C(\langle O(a) = v_1\rangle) = x < 1$, and $C(\langle O(a) = v_1\rangle | a) = y \neq x$. The agent's expectation of the objective value of option a is:

$$\sum_V C(\langle O(a) = v\rangle)v = xv_1 + (1 - x)v_2 = x(v_1 - v_2) + v_2.$$

The V-value of a is:

$$\sum_V C(\langle O(a) = v\rangle | a)v = yv_1 + (1 - y)v_2 = y(v_1 - v_2) + v_2.$$

If Expectationism is true, the subjective value of a relative to C is equal to the expectation of the objective value of a relative to C. Hence:

$$x(v_1 - v_2) + v_2 = y(v_1 - v_2) + v_2.$$

But $v_1 \neq v_2$ implies that $x(v_1 - v_2) + v_2 = y(v_1 - v_2) + v_2$ only if $x = y$, and $x \neq y$. Therefore, EDT, Relevance, and Expectationism cannot all be true. *QED*.

Note that Expectationism, as we have characterized it, is a thesis about numerical values. It says that the numerical representation of the subjective value of an option relative to a credence function (on some arbitrarily chosen scale) must bear a certain arithmetic relation to the numerical representation of the objective value of the option (on that same scale). But one might be concerned with a more general, purely normative thesis: namely,

> **Normative Expectationism**. Agents always subjectively ought to maximize expected objective value.

The proof above does not establish that EDT, Relevance, and Normative Expectationism are inconsistent, but the inconsistency between these three claims can now be proven very simply. Just imagine a case which, relative to some arbitrary determination of a scale, satisfies Relevance, and suppose without loss of generality that $v_1 > v_2$ and $x > y$. Then imagine a choice between options a and b, where a is as described above and b is a lottery with chance $\frac{x+y}{2}$ of realizing an outcome with objective value v_1 and chance $1 - \left(\frac{x+y}{2}\right)$ of realizing an outcome with objective value v_2. Normative Expectationism implies that the agent subjectively ought to realize a, but EDT implies that the agent subjectively ought to realize b.

Appendix B

Here we prove Result #2: that Newcombizability, Dominance, and EDT are jointly inconsistent. Start from a case that witnesses the truth of Newcombizability. Then, according to the agent's credences: $C(\langle O(a_1) = v_1\rangle \vee \langle O(a_1) = v_3\rangle) = 1$, $C(\langle O(a_1) = v_1\rangle | a_1) = x < 1$, $C(\langle O(a_1) = O(a_2) + z\rangle) = 1$, and $C(\langle O(a_1) = v_1\rangle | a_2) = y > x + \frac{z}{v_1 - v_3}$. Then since EDT entails Conditional Expectationism (see n. 22), the V-value of option a_1 equals:

$$\sum_V C(\langle O(a_1) = v\rangle | a_1)v = xv_1 + (1 - x)v_3 = x(v_1 - v_3) + v_3.$$

The V-value of option a_2 equals:

$$\sum_V C(\langle O(a_2) = v\rangle | a_2)v = yv_2 + (1 - y)v_4 = y(v_1 - z) + (1 - y)(v_3 - z) = y(v_1 - v_3) + v_3 - z.$$

And since $y > x + \frac{z}{v_1 - v_3}$,

$$y(v_1 - v_3) + v_3 - z > \left(x + \frac{z}{v_1 - v_3}\right)(v_1 - v_3) + v_3 - z = x(v_1 - v_3) + v_3.$$

Thus, if EDT is true, the subjective value of a_2 relative to C exceeds the subjective value of a_1 relative to C.

But the agent is certain that the objective value of a_1 exceeds the objective value of a_2. So, if Dominance is true, the subjective value of a_1 relative to C exceeds that of a_2 relative to C. Therefore, Newcombizability, EDT, and Dominance cannot all be true. *QED*.

References

Ahmed, A. 2014. *Evidence, Decision and Causality.* Cambridge: Cambridge University Press.

Buchak, L. 2013. *Risk and Rationality*. Oxford: Oxford University Press.

Bradley, R. and H. O. Stefánsson. 2016. Desire, Expectation and Invariance. *Mind* 125: 691-725.

Broome, J. 1991. Desire, Belief and Expectation. *Mind* 100: 265-7.

Dreier, J. 1993. Structures of Normative Theories. *Monist* 76: 22-40.

———. 2011. In Defense of Consequentializing. In Timmons, M. (ed.), *Oxford Studies in Normative Ethics, Vol. 1*. Oxford: Oxford University Press: 97-119.

Hájek, A. 2015. On the Plurality of Lewis's Triviality Results. In Loewer, B. and J. Schaffer (ed.), *A Companion to David Lewis*. Oxford: Blackwell: 425-45.

——— and P. Pettit. 2004. Desire Beyond Belief. *Australasian Journal of Philosophy* 82: 77-92.

Hare, C. 2010. Take the Sugar. *Analysis* 70: 237-47.

———. 2016. Should We Wish Well to All? *Philosophical Review* 125: 451-72.

Horgan, T. 1981. Counterfactuals and Newcomb's Problem. *Journal of Philosophy* 78: 331-56.

Jackson, F. 1991. Decision-Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics* 101: 461-88.

Jeffrey, R. C. 1983. *The Logic of Decision*. Second ed. Chicago: University of Chicago Press.

Joyce, J. 1999. *The Foundations of Causal Decision Theory.* Cambridge: Cambridge University Press.

Kolodny, N. and J. MacFarlane. 2010. Ifs and Oughts. *Journal of Philosophy* 107: 115-43.

Lewis, D. K. 1973. *Counterfactuals*. Oxford: Blackwell.

———. 1979. Counterfactual Dependence and Time's Arrow. *Noûs* 13: 455-76. Reprinted in his *Philosophical Papers Volume II* (1986). Oxford: Oxford University Press.

———. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59: 5-30.

———. 1988. Desire as Belief. *Mind* 97: 323-32.

———. 1996. Desire as Belief II. *Mind* 105: 303-13.

Louise, J. 2004. Relativity of Value and the Consequentialist Umbrella. *Philosophical Quarterly* 54: 518-36.

McDermott, M. 1999. Counterfactuals and Access Points. *Mind* 108: 291-334.

Mackie, J. L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin Books.

Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press.

Oddie, G. 1994. Harmony, Purity, Truth. *Mind* 103: 451-72.

——— and P. Menzies. 1992. An Objectivist's Guide to Subjective Value *Ethics* 102: 512-33.

Parfit, D. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

———. Unpublished. What We Together Do. http://individual.utoronto.ca/stafforini/parfit/parfit_-_what_we_together_do.pdf

Pettigrew, R. 2015. Risk, Rationality, and Expected Utility Theory *Canadian Journal of Philosophy* 45: 796-826.

Portmore, D. W. 2007. Consequentializing Moral Theories. *Pacific Philosophical Quarterly* 88: 39-73.

Price, H. 1989. Defending Desire-as-Belief. *Mind* 98: 119-27.

Regan, D. 1980. *Utilitarianism and Cooperation*. Oxford: Oxford University Press.

Spencer, J. and I. Wells. Forthcoming. Why Take Both Boxes? *Philosophy and Phenomenological Research*.

Thomson, J. J. 2008. *Normativity*. Peru, Illinois: Open Court.

Zimmerman, M. 2008. *Living With Uncertainty*. Cambridge: Cambridge University Press.