Objective Value Is Always Newcombizable

Arif Ahmed
Gonville and Caius College, Cambridge
ama24@cam.ac.uk

Jack Spencer
Massachusetts Institute of Technology
jackspen@mit.edu

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

After some scene-setting (§§1-3), we consider three arguments for our thesis: the argument from Newcomb's Problem (§§4-5), the argument from Expectationism (§§6-7), and the argument from Newcombizability (§8). The first two arguments fail for instructive reasons. But the third succeeds. EDT is incompatible with options having objective values because objective value is always Newcombizable.

What to make of this incompatibility is a matter on which the authors disagree. One is inclined to take it to be a reason for rejecting the claim that options have objective values; the other is inclined to take it to be a reason for rejecting EDT. The paper, itself, takes no stand on this downstream disagreement; it merely argues for the incompatibility.

1. Objective and subjective

*Ought*s come in many flavours — there are moral and epistemic and aesthetic *ought*s — and our arguments apply equally to any. But we'll focus on prudential *ought*s, partly to keep things simple, and partly because it is there that our arguments seem to us to have the most interesting application.

Mineshaft cases have convinced many philosophers of a distinction between an objective and a subjective moral *ought*,[1] and, if we assume that an agent's concerns align with morality, we can use the same cases

[1] See, for example, Jackson (1991), Oddie and Menzies (1992), Parfit (unpublished), and Regan (1980).

Mind, Vol. 00 . 0 . November 2019
doi:10.1093/mind/fzz070

to motivate a distinction between an objective and a subjective prudential *ought*. Consider:

> *Miners.* Ten miners are trapped in shaft A or shaft B, but the agent does not know which. Flood waters threaten the shafts. The agent, who cares only about saving lives, 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 (prudentially) ought the agent to do? And is the agent's uncertainty relevant to what she (prudentially) ought to do? If there is both an objective and a subjective *ought*, 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% confident that they are in shaft B, then the agent subjectively ought to block neither shaft — despite knowing for certain that that is not what she objectively ought to do.

This essay concerns theories of the objective and subjective *ought* that could be defended by a (maximizing) consequentialist. (The structural nature of our argument makes it applicable to any flavour of *ought* for which the objective/subjective distinction can be drawn. But, as we said above, we'll focus on prudential *ought*s.)

If a consequentialist accepts that there is an objective *ought*, then they need to provide some account of it. They need to find some property of options such that: whenever an option is objectively permissible, it is so in virtue of maximizing that property. We'll take *objective value* to be the property of options that features in the consequentialist theory of the objective *ought*, and we'll take rival consequentialist theories of the objective *ought* to be rival hypotheses about what objective value is.

Later on (see §3 and §5), we'll say more about the nature and structure of objective value. For now, what matters is the connection between objective values and the objective *ought*. As we use the terms,

the claim that options have objective values is equivalent to the claim that some consequentialist theory of the objective *ought* is true. In arguing that EDT is inconsistent with options having objective values, we take ourselves to be arguing that EDT is inconsistent with any consequentialist theory of the objective *ought*.[2] To put it another way: EDT is inconsistent with objective consequentialism.

An exactly parallel situation arises on the subjective side. If a consequentialist accepts that there is such a thing as the subjective *ought*, then they need to provide some account of it. They need to find some property of options such that: whenever an option is subjectively permissible, it is so in virtue of maximizing that property. We'll take *subjective value* to be the property of options that features in the consequentialist theory of the subjective *ought*, and we'll take rival consequentialist theories of the subjective *ought* to be rival hypotheses about what subjective value is.

2. Two conceptions of subjective value

EDT is a hypothesis about what subjective value is: subjective value is *evidential expected value*.

We can characterize evidential expected value precisely using four bits of formalism.

First, a set of possible worlds. Let $W = \{w_1, w_2, \ldots, w_m\}$ be the set of *possible worlds* and, for simplicity, assume that W is finite.[3] We'll take *propositions* to be subsets of W, and, where it eases the formalism, we'll ignore the distinction between a world and its singleton.

Second, for any decision situation, we need a set of options, $A = \{a_1, a_2, \ldots, a_n\}$. We'll take options to be propositions that the agent can make true by deciding, and we'll assume that, in any decision situation, options are pairwise exclusive and jointly exhaustive.

Third, we need a specification of the value of each world. Values, like *ought*s, come in many flavours. If we were concerned with moral

[2] In fact, we think EDT is inconsistent with any *consequentializable* theory of the objective *ought* — that is, any theory on which there is some quantity that an agent's objectively permissible options invariably maximize, whether or not the theory treats objective values as conceptually or metaphysically prior to the objective *ought*. Some philosophers conjecture that every (plausible) theory of the objective *ought* is consequentializable; see, for example, Dreier (1993; 2011), Louise (2004), and Portmore (2007). If this is true, then our arguments below establish that EDT is inconsistent with every (plausible) theory of the objective ought.

[3] We also assume $|W| \geq 4$.

*ought*s, we would be concerned with the moral value of a world.[4] Since we focus on prudential *ought*s, we focus on prudential value, where the prudential value of a world is the degree to which the decision-making agent desires that the world be actual.

We assume that prudential value has *ordinal structure*. In other words, we assume a strict weak ordering of possible worlds by their prudential value. Some flavours of value may lack ordinal structure. In fact, prudential value might lack ordinal structure.[5] But EDT — the claim that subjective value of a given flavour is evidential expected value of that flavour — cannot be true unless the relevant flavour of value has at least ordinal structure.[6] And since we are trying to show that EDT is inconsistent with options having objective values, it's legitimate for us to grant assumptions about the structure of value that EDT presupposes.

Having assumed that prudential value has ordinal structure, we introduce numerical representations.[7] Let u be a *world-valuation function* mapping each world to a real number. We say that u is an *ordinally adequate representation* of prudential value just if: $u(w_1) > u(w_2)$ if and only if w_1 is more prudentially valuable than w_2. If u is an ordinally adequate representation of prudential value, then so are all (and only) positive monotone transformations of u.

The fourth thing we need is a credence function. Let C be the agent's credence function: a probability function that represent the agent's confidence in each proposition. In the usual way, let $C(p)$ be the agent's credence in p, and let $C(p|q) = C(pq)/C(q)$, if defined, be the agent's conditional credence in p given q.

[4] The moral value of a world is determined by the morally relevant goings on therein. For example, according to a rather crude hedonism, the moral value of a world is the pleasure net of pain that occurs there.

[5] See, for example, Bader (2017), Bales, Cohen, and Handfield (2014), Chang (2005), Doody (2019), Hare (2010), and Schoenfield (2014).

[6] To anticipate: the evidential expected value of an option (a set of worlds) is a weighted average of the values of those worlds. But a weighted average of values makes no sense unless those values can be weakly ordered.

[7] At least this is so if W is countable. If W is uncountable, and in particular if world-equivalence on the relevant flavour of value partitions W into uncountably many equivalence classes, then some lexicographic ordinal structures are not numerically representable; see Kreps (1988, pp. 24-5). *We* assume finitely many worlds, but this is not strictly necessary: we need only assume that the relevant flavour of value partitions W into countably many *equivalence classes*.

With these four bits of formalism, we can characterize the evidential expected value of option a (relative to credence function C and world-valuation function u):

$$\textit{Evidential Expected Value}: \quad V(a) = \sum_{W} C(w|a)u(w)^{8}.$$

According to EDT, evidential expected value is subjective value. In other words, according to EDT: if C represents an agent's beliefs and u represents prudential value (the agent's desires), then what makes an option subjectively permissible for the agent to choose is the maximization of V (relative to C and u).

The chief rival of EDT is causal decision theory (CDT). 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*.[9] 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 this measure, *uniquely* closest to w_i;[10] (ii) it is *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 $a \Rightarrow w_j$ is the nonbacktracking counterfactual — that world w_j would have been actual had option a been chosen — then $a \Rightarrow w_j$ 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.

[8] 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, p. 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 if value is not 'gunky', as Hájek (2015, p. 438) puts it.

[9] For other characterizations, see, for example, Ahmed (2014, pp. 48-54), Joyce (1999, ch. 5), and Lewis (1981).

[10] No natural similarity metric is quite this discerning: for discussion see Lewis (1973) and Joyce (1999, sect. 5.4). A more realistic proposal: 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.

The causal expected value of option a (relative to C and u) is a function of the agent's credences in these nonbacktracking counterfactuals:

$$\textit{Causal Expected Value}: \; U(a) = \sum_{W} C(\langle a \Rightarrow w \rangle) u(w).$$

According to CDT, subjective value is causal expected value. So according to CDT: if C represents an agent's beliefs and u represents prudential value, then what makes options subjectively permissible for the agent to choose is the maximization of U (relative to C and u).

3. Interval structure[11]

We assumed that prudential value has ordinal structure. But, in fact, proponents of EDT and CDT almost always assume that prudential value has, not just ordinal structure, but *interval structure* (or 'affine structure'). And for good reason: EDT and CDT are plausible accounts of subjective (prudential) value only if (prudential) value has at least interval structure.

Focus on EDT. In principle, one could combine EDT with the view that prudential value has only ordinal structure. But the result is highly unattractive, for it implies that in a wide range of cases there will be no fact of the matter about what an agent is subjectively permitted to do. Consider the world-valuation functions that are ordinally adequate representations of prudential value. If prudential value has only ordinal structure, none of these functions does a better job of representing prudential value than any other. All are adequate representations of prudential value, *simpliciter*. But the ranking of options vis-à-vis evidential expected value might depend on which of these functions we use. Let $A = \{a_i, a_j, \ldots, a_n\}$ be the set of the agent's options; let C be the agent's credence function; and let u_1 and u_2 be two ordinally adequate representations of prudential value. Then it could be the case that a_i uniquely maximizes evidential expected value relative to C and u_1, whereas $a_j \neq a_i$ uniquely maximizes evidential expected value relative to C and u_2. And if this sort of conflict arises, then, according to EDT, there will be no fact of the matter about what the agent is subjectively permitted to do.

[11] Many thanks to a referee for helping us to get clearer on the issues that this section discusses.

Here's an illustration. Suppose the agent cares only about how many smiles there are, and suppose that she is choosing between an option a_i on which (she is certain) there is a 50% chance of there being exactly one smile and a 50% chance of there being 99 smiles, and an option a_j on which (she is certain) there is a 100% chance of there being exactly two smiles. If $n(w)$ is the number of smiles at w, then both of the following functions are ordinally adequate representations of prudential value: $u_1(w) = n(w)$ and $u_2(w) = 1 - 2^{-n(w)}$. But, while a_i uniquely maximizes evidential expected value relative to C and u_1, a_j uniquely maximizes evidential expected value relative to C and u_2. So, if prudential value has only ordinal structure, then, according to EDT, there is no fact of the matter about what the agent is subjectively permitted to do.

Indeed, if prudential value has only ordinal structure, then there is no fact of the matter about what the agent in Miners is subjectively permitted to do. The claim that the agent in Miners is subjectively required to block neither shaft entails that prudential value has more than mere ordinal structure.

If proponents of EDT want to ensure that there is always a fact of the matter about what an agent is subjectively permitted to do (as they should), then they must claim that prudential value has interval structure. To say that prudential value has interval structure is to say that, for any worlds w_1, w_2, w_3, where w_1 and w_3 have unequal prudential value, there is a fact about the ratio of the *difference* in prudential value between w_2 and w_1 to the *difference* in prudential value between w_3 and w_1. If prudential value has interval structure, then a world-valuation function u is an *interval-adequate representation* of prudential value just if: for any possible worlds w_1, w_2, and w_3, where w_1 and w_3 are of unequal prudential value, $\frac{u(w_2)-u(w_1)}{u(w_3)-u(w_1)}$ is equal to the ratio of the difference between the prudential value of w_2 and w_1 to the difference between the prudential value of w_3 and w_1.[12]

[12] Here is a proof that if a proponent of EDT wants to ensure that there is *always* a fact about what an agent is subjectively permitted to do, then they must claim that prudential value has interval structure. Let u_1 and u_2 be ordinally adequate valuation functions on worlds and suppose $\frac{u_1(w_2)-u_1(w_1)}{u_1(w_3)-u_1(w_1)} > \frac{u_2(w_2)-u_2(w_1)}{u_2(w_3)-u_2(w_1)}$, where w_1 and w_3 have unequal prudential value. Since by ordinal adequacy $u_1(w_2) = u_1(w_1)$ if and only if $u_2(w_2) = u_2(w_1)$, it follows that $u_1(w_2) \neq u_1(w_1)$, and we may assume that $u_1(w_3) < u_1(w_2) < u_1(w_1)$ since a parallel proof works for all other cases. Suppose $\frac{u_1(w_2)-u_1(w_1)}{u_1(w_3)-u_1(w_1)} = p$ and let a be a lottery that gives a chance of p to w_3 and a chance of $1-p$ to w_1. Then relative to u_1, EDT regards a as subjectively permissible in a choice between it and $\{w_2\}$; but relative to u_2 it regards a as

Since EDT is plausible only if prudential value has interval structure, we assume, hereafter, that prudential value has interval structure. We do not *need* this assumption. As Appendix C shows, our main argument — the argument from Newcombizability — works even in an ordinal setting. But in the main text, we assume that value has interval structure. So, if u adequately represents prudential value, then so do all (and only) positive linear transformations of u.

Having assumed that prudential value has interval structure, we can pick a unique scale on which to ascribe values to worlds. Pick arbitrary worlds w_0 and w_1 such that w_1 realizes more prudential value than w_0. Then our preferred scale of world-valuation u treats w_1 as its unit and w_0 as its zero. So, for example: $u(w) = 5$ if and only if the difference in prudential value between w and w_0 is five times the difference in prudential value between w_1 and w_0. Since all adequate representations of prudential value agree on ratios of intervals, this ratio will not depend on which world-valuation function we focus upon. The same goes for any flavour of value for which one might want to defend EDT.

Having said a bit about subjective (prudential) values, we turn to objective (prudential) values. As we'll see, although the (prudential) value of a *world* is always a necessary matter, the objective (prudential) values of *options* (that is, sets of worlds) is often both contingent and open to rational doubt. This fact will play an important role in the argument to come.

4. The argument from Newcomb's Problem

Consequentialists who believe in the objective *ought* often assume that the objective value of an option is its *causal value*.

The causal value of an option can be characterized by appeal to the highly discerning causal similarity metric from above. The causal value of option a at world w_i (relative to u) is the value of the a-world that is, by the lights of the highly discerning causal similarity metric, closest

impermissible. So if EDT gives determinate advice then $\frac{u_1(w_2)-u_1(w_1)}{u_1(w_3)-u_1(w_1)} \leq \frac{u_2(w_2)-u_2(w_1)}{u_2(w_3)-u_2(w_1)}$; by a similar argument $\frac{u_1(w_2)-u_1(w_1)}{u_1(w_3)-u_1(w_1)} \geq \frac{u_2(w_2)-u_2(w_1)}{u_2(w_3)-u_2(w_1)}$. Therefore u_1 and u_2 must agree over these interval ratios. For a more sophisticated, axiomatic treatment of the relation between the affine character of a scale and its application to 'weighted averages' see Von Neumann and Morgenstern (1953, pp. 16-27).

to w_i. Thus, for example, if $\langle a \Rightarrow w_j \rangle$ is true at world w_i, then the causal value of a at w_i (relative to u) is $u(w_j)$.[13]

If objective value is causal value, then we can see why the objective value of an *option* (if it is a set of worlds but not a singleton) is both contingent and open to doubt. Let $\langle O(a) = v \rangle$ be a proposition ascribing objective value v to option a. If objective value is causal value, then $\langle O(a) = v \rangle$ is true at w_i if and only if the causally closest a-world to w_i has a u-value of v. Thus $\langle O(a) = v \rangle$ may be true at some worlds but not others, and an agent who is unsure which world is actual may be uncertain whether $\langle O(a) = v \rangle$ is true. For example, if the u-value of a world is just the number of miners that the agent saves at that world, then the objective value of blocking shaft A is 10 at worlds in which the miners are trapped in shaft A and zero at worlds in which the miners are trapped in B, and the agent in Miners, because she is rationally uncertain about where the miners are, is rationally uncertain whether the objective value of blocking shaft A is 10.[14]

If objective value is causal value, then we can establish our thesis — that EDT is inconsistent with options having objective values — via *the argument from Newcomb's Problem.*

EDT and CDT often agree on what maximizes subjective value, but not always. One famous crux is:

> *Newcomb's Problem.* There is a transparent box and an opaque box. The agent, who cares only about money, has two options: take only the opaque box ('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, the opaque box contains $0. If the predictor predicted that the agent would one-box, the opaque box contains $1,000,000. The agent knows all this.

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

[13] See, for example, Moore (1903, p. 149).

[14] Moreover, if objective value is causal value then the objective value of a particular world, that is to say of its singleton, *is* necessary. If $a = \{w\}$, then w is by any measure the closest a-world to any world, so the proposition $\langle O(a) = u(w) \rangle$ is true at every world (and the agent is certain that it is true).

But two-boxing uniquely maximizes U. If we interpret 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 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 equals her credence that she would have received \$1,000,000 had she one-boxed. The U-score of two-boxing is therefore 1,000 greater than that of one-boxing, whatever the U-score of one-boxing is.

The argument from Newcomb's Problem refers to Newcomb's Problem and has two premises:[15]

> *Newcomb Knowledge*: If options have objective values then: an agent facing Newcomb's Problem knows for certain that two-boxing uniquely maximizes objective value;
>
> *Bridge*: If options have objective values then: if an agent's options are a_1, a_2, …, 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 assumption that objective value is causal value. An agent facing Newcomb's Problem knows for certain that the causal value of two-boxing exceeds that of one-boxing. So, if objective value is causal value, Newcomb Knowledge is true.

Bridge is motivated by metaethical considerations.[16] The subjective *ought* arises from the agent's subjective uncertainty about the objective *ought*. When an agent is uncertain about what she objectively ought to do, discrepancies can arise between what an agent subjectively and objectively ought to do. Indeed, as Miners illustrates, when an agent is uncertain about what she objectively ought to do, she can know for certain that what she subjectively ought to do differs from what she objectively ought to do. But if an agent knows for certain what she objectively ought to do, the objective and subjective *ought* must coincide. The following conditional is therefore true (for any flavour of

[15] Spencer and Wells (2019) use something akin to the argument from Newcomb's Problem to defend two-boxing.

[16] Schoenfield (2014) discusses a principle she calls LINK, which resembles, but is stronger than, Bridge. (For semi-critical discussion of LINK, see Doody (2019).) Schoenfield's sympathy for Bridge-like principles is not exceptional. *Many* philosophers accept Bridge. In fact, many accept strictly stronger (alleged) connections between the objective and the subjective *ought*. For example, anyone who, like Parfit (1984, pp. 24-5), equates the subjective value of an option with the agent's expectation of its objective value, is committed to Bridge.

ought): if an agent knows for certain that she objectively ought to choose option a_i, she subjectively ought to choose option a_i.[17] And, given the sort of consequentialism that we are assuming, this is equivalent to Bridge.[18]

One can formulate a principle like Bridge, which applies both to requirements *and* to permissions:

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

We can't see any reason to accept Bridge and reject Bridge Permission. (In our view, both are obviously true.) But since all the cases we discuss fall under Bridge, we focus on Bridge, and set Bridge Permission aside.[19]

[17] 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. But it is not the case that the agent subjectively ought to block some shaft. 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 pairwise exclusive, *blocking some shaft* cannot be an option if *blocking shaft A* is an option.

[18] Bridge must be distinguished from nearby principles.

One 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 principle is false. It might be that an agent, who is uncertain whether p, objectively ought to pay a small sum of money to learn whether p, despite knowing for certain that she would be objectively required not to pay the 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 is known. *We* do not think 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.

While Hare's (2016) example has no force against Bridge, it does raise another difficultly for us, as was pointed out by a helpful referee. We discuss that difficulty in n. 33.

[19] One might think opaque sweetening cases (*cf.* Hare 2010) threaten Bridge Permission, but they don't. We are operating in a setting where values of worlds form a strict weak order, and opaque sweetening cases cannot arise in this setting.

The Argument from Newcomb's Problem is as follows. If EDT is true, one-boxing uniquely maximizes subjective value. But, by Newcomb Knowledge, if options have objective values, then the agent knows for certain that two-boxing uniquely maximizes objective value. Therefore, by Bridge, if options have objective value, then two-boxing uniquely maximizes subjective value. Therefore, if EDT is true, then options do not have objective values.

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

5. 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 EDTists grant that objective value is causal value? After all, there are *many* conceptions of objective value (where by 'conception of objective value' we mean a hypothesis 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 $\{w_j, w_k, \ldots, w_m\}$ 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 (relative to u) is some weighted average of $u(w_j)$, $u(w_k)$, …, $u(w_m)$, the values of the a-worlds that are, by the lights of the similarity metric, most similar

If we want to operate in a setting where opaque sweetening cases *can* arise, then we might represent an agent's desires, not with a single world-valuation function, but with a set of them, which we could call a *representor*. Even here, there are Bridge-like principles that we accept. For example:

> If options have objective values then: if an agent knows for certain that option a_i maximizes objective value relative to each member of the representor, then the agent is subjectively permitted to choose a_i.

We are more doubtful about principles covering cases in which there is disagreement in the representor. For example:

> If options have objective values then: if an agent knows for certain that option a_i uniquely maximizes objective value relative to at least one member of the representor, then the agent is subjectively permitted to choose a_i.

Fortunately, our arguments do not require any stand on such principles. For more on incommensurability and opaque sweetening, see, for example, Bader (2017), Bales, Cohen, and Handfield (2014), Chang (2005), Doody (2019), and Schoenfield (2014).

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 just the value of the a-world that is, by the lights of the similarity metric, uniquely most similar to w_i.

Call a similarity metric *m trivial* if for any option a and any worlds w_1 and w_2, the most m-similar a-worlds to w_1 are the same as the most m-similar a-worlds to w_2. If the similarity metric underlying objective value is non-trivial and some option a is true at more than one world, then there is almost certainly some proposition $\langle O(a) = v \rangle$ that is contingent and open to rational doubt. (The argument for this is the same as at §4 regarding the causal metric.)

One conception of objective value that falls within this natural class is causal value. But there are many others, and many are uniquely maximized by *one-boxing* in Newcomb's Problem.[20]

For instance, consider Horgan's defence 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 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]. (Horgan 1981, p. 336, emphases in original)

Horgan does not give details, but we can devise a similarity metric on possible worlds meeting his desiderata. One way to do it would be to

[20] The full class of conceptions of objective value extends beyond this natural family. In its broadest sense, a conception of objective value is *any* function taking each (possible world, proposition) pair to a real number. Our result extends to this broader class.

graft Horgan's 'top priority' onto something like Lewis's (1979) lexicographic criteria for relative similarity, insisting that what counts most for closeness of a world w is whether the predictor's accuracy at w with regard to the agent's choice matches the predictor's actual accuracy on that question.[21] We then use the Horgan metric to define a conception of objective value, which we 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 $u(w_j)$.

Often, the Horgan value of an option is its causal value. For instance, they coincide in Miners.

But in Newcomb's Problem, they diverge. The causal metric holds fixed the contents of the opaque box but not the predictor's correctness. Whatever is in the opaque box at the two-boxing world that is causally closest to actuality 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

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

(1) $w >_1 w'$ if and only if: 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) $w >_2 w'$ if and only if: there are big, widespread and diverse violations of the laws of w_i at w' and not at w;

(3) $w >_3 w'$ if and only iff: the spatio-temporal region throughout which perfect match over particular facts with w_i prevails is larger at w than at w'.

(4) $w >_4 w'$ if and only if: there are small, localized, simple violations of the laws of w_i at w' and not at w;

(5) $w >_5 w'$ if and only if: 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) $\{n|w_k >_n w_j\} = \varnothing$ and $\{n|w_j >_n w_k\} \neq \varnothing$; or (ii) $\min\{n|w_k >_n w_j\} > \min\{n|w_j >_n w_k\}$.

Criteria 2)-5) follow Lewis (1979, pp. 47-8), minus Lewis's hedging over 5). Taken jointly as analysing of the 'standard resolution' closeness in natural language, 2)-5) are implausible. (See, for example, McDermott 1999.) *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.

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. Either way, 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 is *not* 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.[22]

The argument from Newcomb's Problem is therefore too quick. It establishes *something*: since Bridge is true, it establishes that EDT is inconsistent with any conception of objective value that validates Newcomb Knowledge. But most conceptions of objective value do not validate 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.

6. 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.

As in §2, choose some similarity metric m, so that the objective-value-assigning proposition $\langle O(a) = v \rangle$ is true at exactly those worlds to which the m-closest a-world w is such that $u(w) = v$ relative to our chosen unit and zero for u. We can suppose that the agent knows the

[22] Suppose the prediction is at t_1, and consider 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 that have positive chance at t_0 at w_i, conditional on a. According to the proposal, the objective value of a at 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 $(1{,}000)(0.99) + (0.01)(1{,}001{,}000) = 11{,}000$.

scale (that is, with a defined unit and zero as at §3), and thus knows the objective value of every option at every world. But since the agent may be uncertain about which world is *actual*, her credence in $\langle O(a) = v\rangle$ may be strictly between 0 and 1.

The formal statement of Expectationism is as follows:

Expectationism: The subjective value of a (relative to C and u) is $\sum_v vC(\langle O(a) = v\rangle)$.

Expectationism is widely accepted. Often it's just assumed;[23] but it has some explicit defences. For example, some try to defend Expectationism by claiming that the subjective value of an option should equal 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.[24]

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 its causal value.[25] 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, we get *Horgan decision theory* (HDT) — the view that the subjective value of an option is the agent's expectation of its Horgan value. Whereas two-boxing uniquely maximizes the agent's expectation of causal value (that is, causal expected value), one-boxing uniquely maximizes the agent's expectation of Horgan value.

But, while Expectationism is not hostile to one-boxers, it *is* hostile to EDT — as we now argue.

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

[23] See, for example, Parfit (1984, p. 25).

[24] See, for example, Oddie and Menzies (1994) and Pettigrew (2015).

[25] *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 $\sum_W C(w)O_w(a) = \sum_{W_1} C(w)u(w_1) + \ldots + \sum_{W_n} C(w)u(w_n) = C(\langle a \Rightarrow w_1\rangle)u(w_1) + \ldots + C(\langle a \Rightarrow w_n\rangle)u(w_n) = \sum_W C(\langle a \Rightarrow w\rangle)u(w) = U(a)$.

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, consider a variation on Miners. On this variation, the agent remembers the miners' telling her which shaft they would be in, but cannot consciously recall which. As before, she is 50% confident that they are in shaft A and 50% confident that they are in shaft B. But she (reasonably) thinks there is a nonzero chance that her unconscious memory will influence her choice if she chooses to block a shaft. 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, in this variant of Miners, 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.[26]

[26] In saying this, we set 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 mind-independent world but also on the beliefs of the agent. We agree with Hájek and Pettit that indexical value concepts evade Lewis's arguments. They may also violate Relevance. For instance, if the indexical value of an option a is just $V(a)$, then, if conditionalization can change the evidential expected value of the tautology (see Bradley and Stefánsson 2016, pp. 699-702), Relevance as applied to indexical value fails.

But we deny that any indexical value that violates Relevance is a plausible notion of *objective* value. 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, whatever the agent thinks. More generally, we require that for any flavour of objective value, it is *possible* for the objective value of an option to depend on circumstances (i) about which the agent is not sure and (ii) to which the agent's choice of

Now we can prove:

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

The 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 over the subjective value of an option. For instance, in the variant on Miners, Expectationism implies (given a natural choice of zero and unit worlds) 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 values because, if options have objective values, Relevance and Expectationism are both true.

The argument from Expectationism has an important advantage over the argument from Newcomb's Problem. The argument from Newcomb's Problem is too narrow. It shows that EDT is inconsistent with any conception of objective value that validates Newcomb Knowledge, but it's silent about conceptions of objective value that do not. 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.

7. 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. CDTists and HDTists agree that subjective value is expected objective value and disagree about what

option is evidentially relevant. (For more on DAB see n. 31; for a related concern about Relevance, see n. 33.)

objective value is, but EDTists do not share this agreement. *No* conception of objective value stands to EDT as causal value stands to CDT.

But Expectationism can be questioned. If options have objective and subjective values, then there must be *some* well-behaved, intimate relationship between the subjective value of an option 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 following principle:

> *Certain Reflection*: If $C(\langle O(a) = v\rangle) = 1$, then the subjective value of option a (relative to C and u) equals v.

Certain Reflection is strictly weaker than Expectationism: there are views that verify Certain Reflection while falsifying Expectationism.

First consider *Maximin* — the view that the subjective value of a is the least v such that the agent assigns nonzero credence to $\langle O(a) = v\rangle$. Maximin satisfies Certain Reflection, but 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 weight to some possibilities than to others depending on their rank and not only on their probability.[27] On most such weightings, the view that subjective value is the risk-adjusted expectation of objective value falsifies Expectationism, but 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 which you can weight according to their rank.[28]

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

[27] Buchak (2013, ch. 2).

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

Conditional Expectationism: The subjective value of a (relative to C and u) is $\sum_v vC(\langle O(a) = v\rangle|a)$.[29]

Conditional Expectationism entails Certain Reflection. And, given Certain Reflection, EDT entails Conditional Expectationism.[30]

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, they 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 favour 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 EDTists must accept Expectationism is thus unjustified. (Even among opponents of EDT, Expectationism is not common ground.) So, without some independent argument that EDTists must accept Expectationism, the argument from Expectationism fails to establish our thesis.[31]

[29] Oddie (1994, p. 460). Also see Broome's (1991) discussion of Conditional Expectationism (which he calls Desire-as-Expectation).

[30] *Proof*: Let w_i be any world at which a and $\langle O(a) = v\rangle$ both hold. Let C concentrate all 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 $u(w_i) = v$. Hence EDT entails Conditional Expectationism, since it follows that for any C, $\sum_{w\in W} C(w|a)u(w) = \sum_{w\in\langle O(a)=v_i\rangle\cap a} C(w|a)u(w) + \ldots = \sum_{w\in\langle O(a)=v_i\rangle} C(w|a)v_1 + \ldots = \sum_v vC(\langle O(a) = v\rangle|a)$.

[31] We should briefly relate this discussion to Lewis's (1986, 1988) argument against the anti-Humean 'Desire as Belief' (DAB) thesis, which the argument from Expectationism obviously resembles. 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 easily extends to cover a graded notion, like the argument from Expectationism.) Given Lewis's assumption that $V(A|A) = V(A)$ (see Bradley and Stefánsson 2016, pp. 699-702 for dissent) 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, p. 309). One way for anti-Humeans to resist Lewis (Price 1989, p. 122) would be to reformulate their thesis as 'Desire as *Conditional* Belief' (DACB): $V(A) = C(\dot{A}|A)$. This gives no traction to Lewis's argument, for the same reason that Conditional Expectationism gives none to the argument from Expectationism. In response, Lewis argues (1996, pp. 310-1) that DACB implies *desire by necessity*: it is committed to a proposition G which the agent desires true *whatever* her

8. The argument from Newcombizability

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

It starts from a principle that is stronger than Bridge, but no less obviously true. As we said, Bridge seems undeniable. If options have objective values then: if an agent's options are a_1, a_2, …, a_n, and the agent knows for certain that a_i uniquely maximizes objective value, then a_i uniquely maximizes subjective value. But the subjective values of options supervene on the agent's credences and the values of worlds. We therefore can weaken the antecedent of Bridge, appealing to any of the agent's certainties, 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, …, a_n, and the agent is certain that option a_i uniquely maximizes objective value, then a_i uniquely maximizes subjective value.

Dominance is helpfully compared to 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 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 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.[32] But

credences. We are not clear why this threatens DACB. Maybe the idea is that the anti-Humean thesis is a descriptive psychological thesis about a person's desires, and it is false as a matter of fact that there is anything that everyone desires (cf. Lewis 1996, pp. 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.

[32] Proof that maximin implies dominance: let $\underline{a} = \min\{x | C(\langle O(a) = x\rangle) > 0\}$. If $\underline{a} \leq \underline{b}$ then $C(\langle O(a) \leq \underline{b}\rangle) > 0$, 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. 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 that of option b equals $\sum_W C(w)(O_w(a) - O_w(b))$. Since

Dominance is hostile to EDT. There is a generalization of Relevance — a principle 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, although 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 (as measured on the scale determined by our choice of unit and zero point), the agent 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:

$$\text{(i)} \quad C\big(\langle O(a_1) = v_1\rangle \vee \langle O(a_1) = v_3\rangle\big) = 1,$$

$$\text{(ii)} \quad C(\langle O(a_1) = v_1\rangle | a_1) = x < 1,$$

$$\text{(iii)} \quad C(\langle O(a_1) = O(a_2) + z\rangle) = 1, \textit{ and}$$

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

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 that 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:

$$\text{(i)} \quad C\big(\langle O(a_1) = v_1\rangle \vee \langle O(a_1) = v_3\rangle\big) = 1,$$

---

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 by Expectationism entails that, relative to C, the subjective value of a exceeds that of b. Proof that Risk-adjusted Expectationism implies dominance: see Buchak (2013, pp. 245-6). 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 this weakened dominance principle: if an agent's options are a_1, a_2, …, a_n, and the agent is certain that 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.

(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) $C(\langle O(a_1) = v_1\rangle | a_2) = y > x + \frac{z}{v_1 - v_3}$.[33]

As far as we can see, *almost every* conception of objective value is Newcombizable.[34] Almost every conception of objective value entails that there are cases in which the objective values of two options are completely independent of the agent's joint credence distribution over (a) which option she will take and (b) what the objective values of the options are. And given any such case, there is a general recipe

[33] Hare's 2016 case challenges the following strong principle:

Independence Assumption: Any distribution of objective values over options at worlds is consistent with any distribution of credences over those worlds.

Hare's case might convince you (1∗) that killing one to save five is objectively permissible when it is unknown which of the six is being sacrificed to save the others, and (2∗) that killing one to save five is objectively impermissible when the identity of the one being sacrificed is known by the decision-making agent. If (1∗) and (2∗) are both true, and consequentialism is true, then the Independence Assumption is false — since then a change in the agent's credences precipitates a change in the objective values of her options.

As it happens, we do not think that (1∗) and (2∗) are both true. Moreover, even when we *suppose* that (1∗) and (2∗) are both true, we are inclined to accept some version of the Independence Assumption restricted to goods that, unlike human lives, are fungible. For example, the objective value of burning one dollar bill to receive five others is *not* sensitive to whether the agent knows the identity of the bill being burnt.

But even if we were convinced that the Independence Assumption failed for every good, we still would think that every remotely plausible conception of objective value satisfies Relevance, Baseline Relevance, and Newcombizability. If Hare cases establish anything about the credence-dependence of objective values, they establish that the objective values of options can be sensitive to the agent's credence (or knowledge) of *whom* is being benefited and harmed. But the cases that motivate Relevance, Baseline Relevance, and Newcombizability do not involve this sort of identity uncertainty.

[34] Why 'almost'? (i) We assume 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 some such distribution of objective values of options is consistent with any joint distribution of credences across options and their objective values. This seems uncontentious to us, since denying it would impose an exceptionally strong indexicalism (see n. 26), on which the fact that two options have different objective values *by itself* excludes the agent from having certain *credences* concerning which option she will realize and what the objective values of the options are. That would be a far stronger indexicalism than any endorsed by Hájek and Pettit. (ii) is also consistent with any plausible response to Hare-type cases, because the propositions that it concerns, specifying which option is realized and what the objective values of the options are, lack the identity-directed character that Hare's examples exploit: see n. 33.

for Newcombizing. Let O be any conception of objective value. If there are options a_1 and a_2, then there are propositions

Table 1

	S_1	S_2
a_1	v_1	v_3
a_2	$v_2 = v_1 - z$	$v_4 = v_3 - z$

$S_1 =_{def.} \langle O(a_1) = v_1 \wedge O(a_2) = v_1 - z\rangle$ and $S_2 =_{def.} \langle O(a_1) = v_3 \wedge O(a_2) = v_3 - z\rangle$. We can therefore construct a decision problem with the following payoffs:

Assuming that the objective values of these options are completely independent of the agent's joint credence distribution as just described, we can 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 $C(a_2 s_2) = \big(1 - y\big)(1 - k)$. It is easy to check that C satisfies all of clauses (i)-(iv) in the Newcombizability condition. If one needs a backstory, 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.[35]

[35] For illustration: Horgan value is Newcombizable. Suppose you have a choice between taking box 1 and taking box 2, *both* boxes being opaque. Taking box 1 releases 2 units of welfare; taking box 2 does nothing. A Newcombian predictor has predicted your choice; and if you outwit the predictor, you get 8 bonus units of welfare. Moreover, the predictor is somewhat better at predicting people who choose box 1 (success rate 0.625) than at predicting people who choose box 2 (success rate 0.1875) – not because you are antecedently and robustly confident that the prediction is that you 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 Spencer and Wells (2019, pp. 35-6).)

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

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

> *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.

As we saw in §7, the argument from Expectationism can be resisted by a 'conditionalizing' manoeuvre: one can reject Expectationism in favour of Conditional Expectationism. But no analogous conditionalizing manoeuvre can resist the argument from Newcombizability.

To see this, we'll consider a few proposals. We should emphasize that they all either reject or appear to reject Dominance, and we think Dominance is unassailable. Our reason for discussing these three apparent alternatives to Dominance is to bring out the disanalogy between the argument from Newcombizability and the argument from Expectationism, since the latter *can* be resisted by a conditionalizing manoeuvre.

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

> *Conditional Dominance 1*: If options have objective values then: if an agent's options are a_1, a_2, …, 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.[36]

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 she 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.

[36] Formally, we can write the consequent of Conditional Dominance 1: $C\big(\wedge_{k\neq i}\langle O(a_i) > O(a_k)\rangle | a_i\big) = 1 \rightarrow \wedge_{k\neq i}\sigma(a_i) > \sigma(a_k)$, where σ represents subjective value.

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

Conditional Dominance 2: If options have objective values then: if an agent's options are $a_1, a_2, \ldots, a_n$, 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.[37]

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).

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 would then 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 $a_1, a_2, \ldots, a_n$, 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.[38]

But this just rearranges the deckchairs. For any objective value concept O, we'll say that *the O-value of an option were it to be realized* is the O^**-value* of that option. Then it is just as plausible that O^* satisfies our sufficient condition for Newcombizability as that O itself does. That is: there are cases where the O^*-values of two options are completely independent of the agent's joint credence distribution over (a) which option she will take and (b) what the O^*-values of the options are. If this is so, we can similarly Newcombize O^*-value.[39]

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

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

[39] In slightly more detail: suppose $O^*(a) = n \equiv_{def.} a \Rightarrow O(a) = n$ for some selected counterfactual operator $\Rightarrow$. If there are options a_1 and a_2, then there are propositions $S_1 =_{def.} (a_1 \Rightarrow \langle O(a_1) = v_1\rangle) \wedge (a_2 \Rightarrow \langle O(a_2) = v_1 - z\rangle)$ and $S_2 =_{def.} \big(a_1 \Rightarrow \langle O(a_1) = v_3\rangle\big) \wedge (a_2 \Rightarrow \langle O(a_2) = v_3 - z\rangle)$. We can therefore construct a decision problem with payoffs as in Table 1 and a credence function C on the atoms $\{a_i S_j\}_{i,j=1,2}$ as described in the main argument. A simple example: suppose that O is causal value, that the highly discerning causal similarity

All this having been said, our own view is that Dominance, like Certain Reflection, is unassailable. 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 artefacts of the agent's uncertainty about objective values.

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. Opponents of EDT repeatedly 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 in the opaque box at the causally closest one-boxing world is likewise 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 that 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 is 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 EDTist'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.

metric (§2) is strongly centred, and that the O^*-value of a at w is just the O-value of a at the closest a-world to w according to some highly discerning similarity metric. Then the proposition $O^*(a) = n$ just is the proposition $O(a) = n$, so any credence function that Newcombizes O-value also Newcombizes O^*-value.

9. Conclusion

Past this point, the authors part ways. We agree that EDT is inconsistent with options having objective value, but not about how to respond to this fact. One of us is inclined to take the inconsistency to amount to a metaethical refutation of EDT. The other 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.[40]

Appendix A

Here we prove Result #1: Relevance, Expectationism, and EDT are jointly inconsistent. Start from a 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 a is:

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

As we say in §7, if options have both objective and subjective values, Certain Reflection holds. And as we point out in note 30, given Certain Reflection, EDT entails Conditional Expectationism:

$$\begin{aligned}\sum_W C(w|a)u(w) &= \sum_{w\in\langle O(a)=v_1\rangle \cap a} C(w|a)u(w) + \ldots \\ &= \sum_{w\in\langle O(a)=v_1\rangle} C(w|a)v_1 + \ldots = \sum_V C(\langle O(a) = v\rangle|a).\end{aligned}$$

[40] The authors wish to thank two anonymous referees for helpful feedback and encouragement. Arif Ahmed delivered some of this material at a conference held in Cambridge in March 2019 as part of Prof. John Divers's project on Non-Categorical Thought, and he wishes to thank his audience on that occasion, particularly Simon Blackburn, Ruth Byrne, Attila Csordas, John Divers, Richard Holton, and Shyane Siriwardena. He also wishes to thank Wolfgang Schwarz and Robert Stalnaker for very helpful comments on other occasions. He wrote much of his contribution whilst a visitor at the Department of Linguistics and Philosophy at MIT and wishes to thank that institution for its hospitality. He also wishes to thank the Leverhulme Trust for its support via Research Fellowship RF-2018-231\10 and the Effective Altruism Foundation for its support via Research Grant G103268. Jack Spencer wishes to thank, in addition to those above, David Builes, Kevin Dorst, Caspar Hare, Daniel Muñoz, Agustín Rayo, and Ian Wells. The authors are listed in alphabetical order and made equal contributions to the paper.

Hence the *V*-value of *a* is:

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

If Expectationism is true, the subjective value of *a* relative to *C* equals 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*.

One might be concerned, not with Expectationism, but 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. 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 she subjectively ought to realize *b*.

Appendix B

Here we prove Result #2: Newcombizability, Dominance, and EDT are jointly inconsistent. Start from a case that witnesses the truth of Newcombizability. Then, according to the agent's credences: $C\big(\langle O(a_1) = v_1\rangle \vee \langle O(a_1) = v_3\rangle\big) = 1$, $C(\langle O(a_1) = v_1\rangle|a_1) = x < 1$, $C(\langle O(a_1) = O(a_2) + z\rangle) = 1$, 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 (n. 30), the *V*-value of option a_1 equals:

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

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 that 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*.

Appendix C

In §3 we asserted that if a flavour of value has only ordinal structure, then EDT is inconsistent with options having objective values of that flavour. Here we justify that claim.

We want to show that, even in an ordinal setting, if options have objective values, then EDT violates a principle akin to Dominance. If the principle akin to Dominance is true, then we will have shown that, even in an ordinal setting, EDT is incompatible with options having objective values.

To show this *properly* requires some formalism. We give the formal argument below. But the idea is simple and can be illustrated by an example, so we'll start with that.

Suppose that (prudential) value has only ordinal structure, and suppose that options have objective (prudential) values.

Suppose that we have four goods, A, B, C, and D. In terms of value, A > B > C > D. But since we are in a purely ordinal setting, it does not make sense to ask whether the difference in value between A and B is more or less than the difference in value between B and C or C and D.

Suppose an agent is deciding between two envelopes. The first envelope contains either A or C; the second contains either B or D. The item in the chosen envelope will be preserved; the item in the

unchosen envelope will be destroyed. Which items are contained in the envelopes depends on a prediction made yesterday by a reliable predictor. If the predictor predicted that the agent would choose the first envelope, the first envelope contains C and the second contains D. If the predictor predicted that the agent would choose the second envelope, the first envelope contains A and the second contains B.

Since the agent knows all this, the agent knows for certain that choosing the first envelope uniquely maximizes objective value for *any* ordinally adequate valuation function in worlds. At least, the agent knows this if objective value is causal value.

But EDT does not recommend choosing the first envelope. In fact, EDT goes silent. Whether choosing the first envelope or the second envelope maximizes evidential expected value depends on which of the ordinally adequate representations of (prudential) value we choose. Suppose the agent is 80% confident that the prediction is accurate. Then, relative to an ordinally adequate world-valuation function that assigns 10 to A,[41] 5 to B, 4 to C, and 0 to D, choosing the first envelope uniquely maximizes evidential expected value. But relative to an ordinally adequate world-valuation function that assigns (say) 100 to A, 99 to B, 1 to C, and 0 to D, EDT recommends choosing the second envelope. Neither verdict is more 'correct', so EDT gives no determinate advice.

In giving no verdict, EDT violates the following principle, which we think is as obviously true as Dominance:

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

What this example illustrates is that, even in an ordinal setting, if options have objective values, then EDT conflicts with true principles concerning the relationship between objective values and subjective values. And this is why, even in an ordinal setting, EDT conflicts with options having objective values.

This illustration assumes a particular conception of objective value, that is, causal value. But the argument can be formalized and generalized, as follows:

1. Suppose that objective value of worlds has ordinal structure but not interval structure. That is: say that two functions f and g from

[41] That is: assigns 10 to the closest worlds where the agent gets A.

worlds to numbers are *ordinally equivalent* if and only if $\forall w, w'\big(f(w) > f(w') \leftrightarrow g(w) > g(w')\big)$. Then to say that value has ordinal and not interval structure is to say that any f represents values of worlds if and only if any g that is ordinally equivalent to it represents values of worlds. Let Φ be the set of all functions that do represent values of worlds.

2. Let a *distance-measure* be a symmetric function m from world-world pairs to numbers such that for any w and any proposition A there is a unique A-world w' that minimizes $m(w, x)$ for all A-worlds x. The *generalized objective value function for options* is the function O from (world, proposition, objective world value function, distance-measure) quadruples to numbers such that for any value-representing function for worlds f, for any $w, A, m : O\big(w, A, f, m\big) = f(w')$, where w' is the m-closest A-world to w. Then for each distance-measure M, the *objective value concept* Ω_M is the set of all functions ω from world-proposition pairs to numbers such that for some f that represents the values of worlds, for any A and $w, \omega(w, A) = O\big(w, A, f, M\big)$. For instance, causal value is the objective value concept C:

$$\omega \in C \leftrightarrow \exists f \in \Phi : \forall A \forall w : \omega(A, w) = f(\wedge w' : w \in [A \Rightarrow w'])$$

where the counterfactual operator $\Rightarrow$ is based on the highly discerning causal similarity metric introduced at §2. For any objective value concept Ω, let us say that A *is determinately* Ω*-objectively better than* B *at* w if and only if $\omega \in \Omega \rightarrow \omega(A, w) > \omega(B, w)$.

3. A *subjective value function for options* is any function S from (proposition, credence function, world-valuation function) triples to numbers. A *subjective value concept* Σ is a set of functions σ from (proposition, credence function) pairs to numbers such that for some subjective value function for options S, for some objective value-representing function for worlds f, for any $A, Cr : \sigma(A, Cr) = S\big(A, Cr, f\big)$. For instance, evidential decision theory is a subjective value concept EDT, where:

$$\sigma \in EDT \leftrightarrow \exists f \in \Phi : \forall A \forall Cr : \sigma(A, Cr) = \sum_w f(w) Cr(w|A)$$

For any subjective value concept Σ and any credence function *Cr*, let us say that *Cr reckons A to be determinately* Σ*-subjectively better than B* if and only if $\sigma \in \Sigma \rightarrow \sigma(A, Cr) > \sigma(B, Cr)$.

4. The following principle then seems a plausible joint constraint on any objective and subjective value concepts Ω and Σ:

> *Determinate dominance* (DD): If options have objective value then: *if Cr* is certain that *A* is determinately ω-objectively better than any alternative at @, *then C* reckons *A* to be determinately Σ-subjectively better than any alternative.

Less formally, this is saying that if you are *certain* that for *any* acceptable way of measuring the objective value of *worlds*, *A* comes out as having the uniquely highest objective value amongst the available *options*, then for any acceptable way of measuring the objective value of worlds, *A* is what you subjectively ought to do. But if objective value for worlds has ordinal but not interval structure, then EDT violates DD, as this example shows.

> Let Ω_M be an objective value concept and let f be any function that represents objective value of worlds. Let the agent be certain that one of four worlds is actual: w_1, w_2, w_3, w_4 where $f(w_1) > f(w_2) > f(w_3) > f(w_4)$. Let the agent's alternatives be the propositions $A = \{w_1, w_3\}$ and $B = \{w_2, w_4\}$. And let the distance-measure M underlying Ω_M, and the agent's credence function *Cr*, satisfy the conditions in this table. (Ignore the rightmost column for now.)

Table 2

World *w*	*Cr(w)*	*M*-closest *A*-world to *w*	*M*-closest *B*-world to *w*	*f(w)*
w_1	0.1	w_1	w_2	100
w_2	0.4	w_1	w_2	99
w_3	0.4	w_3	w_4	1
w_4	0.1	w_3	w_4	0

> First, we show that *Cr* is certain that *A* is determinately Ω_M-objectively better than any alternative at @. For any f that represents the values of worlds we have $f(w_1) > f(w_2)$, so the *M*-closest *A*-world to @ is objectively better than the *M*-closest *B*-world to @ if @$\in \{w_1, w_2\}$. Similarly, for any f that represents the

values of worlds we have $f(w_3) > f(w_4)$, so the M-closest A-world to @ is objectively better than the M-closest B-world to @ if $@ \in \{w_3, w_4\}$. Since Cr is certain that one of w_1, w_2, w_3, w_4 is actual, it is certain that f scores the M-closest A-world to @ objectively better than the M-closest B-world to @, for any f that represents the values of worlds. Therefore Cr is certain that A is determinately Ω_M-objectively better than any alternative at @.

Now we show that Cr does not reckon A to be determinately EDT-subjectively better than any alternative. To show this we just need an f that represents the values of worlds such that $V(A) \leq V(B)$, where $V(X) = \sum_w f(w)Cr(w|X)$. Let f be as defined in the right-most column of the table. Clearly, $f(w_1) > f(w_2) > f(w_3) > f(w_4)$. Since we suppose only ordinal structure on values of worlds f represents the values of these worlds. But then $V(A) = 20.8$ and $V(B) = 79.2$.

5. We find determinate dominance overwhelmingly plausible. Therefore, if objective value of worlds has ordinal but not interval structure then one of the following is true:

   (a) If EDT is a true theory of subjective rationality, then options do not have objective value;

   (b) The true theory of objective value of options is based on a closeness measure that rules out a closeness-structure like that in the table.

   In our view, (b) is desperate. It is hard to see why the objective-value-relevant notion of closeness should rule out the possibility that there are propositions A and B such that the closest A-world to any world is objectively better than the closest B-world to it, but some B-worlds are objectively better than some A-worlds, which is all that the case requires. Certainly, both objective causal value and objective Horgan value allow these possibilities, as realized in their respective Newcombizations. We conclude that (a) is true. If objective values of worlds have ordinal but not interval structure then EDT rules out objective values for options.

References

Ahmed, Arif 2014, *Evidence, Decision and Causality* (Cambridge: Cambridge University Press).

Bader, Ralf 2017, 'Stochastic Dominance and Opaque Sweetening', *Australasian Journal of Philosophy* 96: 498–36.
Bales, Adam, Daniel Cohen, and Toby Handfield 2014, 'Decision Theory for Agents with Incomplete Preferences', *Australasian Journal of Philosophy* 92: 453–70.
Bradley, Richard and H. Orri Stefánsson 2016, 'Desire, Expectation and Invariance', *Mind* 125: 691–725.
Broome, John 1991, 'Desire, Belief and Expectation', *Mind* 100: 265–7.
Buchak, Lara 2013, *Risk and Rationality* (Oxford: Oxford University Press).
Chang, Ruth 2005, 'Parity, Interval Value, and Choice', *Ethics* 115: 331–50.
Doody, Ryan 2019, 'Parity, Prospects, and Predominance', *Philosophical Studies* 176: 1077–95.
Dreier, James 1992, 'Structures of Normative Theories', *Monist* 76: 22–40.
—— 2011, 'In Defense of Consequentializing', in Mark Timmons (ed.), *Oxford Studies in Normative Ethics, Vol. 1* (Oxford: Oxford University Press): 97–119.
Hájek, Alan 2015, 'On the Plurality of Lewis's Triviality Results', in Barry Loewer and Jonathan Schaffer (eds.), *A Companion to David Lewis* (Oxford: Blackwell): 425–45.
Hájek, Alan and Phillip Pettit 2004, 'Desire Beyond Belief', *Australasian Journal of Philosophy* 82: 77–92.
Hare, Caspar 2010, 'Take the Sugar', *Analysis* 70: 237–47.
—— 2016, 'Should We Wish Well to All?', *Philosophical Review* 125: 451–72.
Horgan, Terence 1981, 'Counterfactuals and Newcomb's Problem', *Journal of Philosophy* 78: 331–56.
Jackson, Frank 1991, 'Decision-Theoretic Consequentialism and the Nearest and Dearest Objection', *Ethics* 101: 461–88.
Jeffrey, Richard C. 1983, *The Logic of Decision*, Second Edition (Chicago: University of Chicago Press).
Joyce, James M. 1999, *The Foundations of Causal Decision Theory* (Cambridge: Cambridge University Press).
Kreps, David M. 1988, *Notes on the Theory of Choice* (Boulder: Westview Press).
Lewis, David 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, Jennie 2004, 'Relativity of Value and the Consequentialist Umbrella', *Philosophical Quarterly* 54: 518–36.
McDermott, Michael 1999, 'Counterfactuals and Access Points', *Mind* 108: 291–334.
Moore, G. E. 1903, *Principia Ethica* (Cambridge: Cambridge University Press).
Oddie, Graham 1994, 'Harmony, Purity, Truth', *Mind* 103: 451–72.
Oddie, Graham and Peter Menzies 1992, 'An Objectivist's Guide to Subjective Value', *Ethics* 102: 512–33.
Parfit, Derek 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, Richard 2015, 'Risk, Rationality, and Expected Utility Theory', *Canadian Journal of Philosophy* 45: 796–826.
Portmore, Douglas W. 2007, 'Consequentializing Moral Theories', *Pacific Philosophical Quarterly* 88: 39–73.
Price, Huw 1989, 'Defending Desire-as-Belief', *Mind* 98: 119–27.
Regan, Donald H. 1980, *Utilitarianism and Co-operation* (Oxford: Oxford University Press).
Schoenfield, Miriam 2014, 'Decision Making in the Face of Parity', *Philosophical Perspectives* 28: 263–77.
Spencer, Jack and Ian Wells 2019, 'Why Take Both Boxes?', *Philosophy and Phenomenological Research* 99: 27–48.
Von Neumann, John and Oskar Morgenstern 1953, *Theory of Games and Economic Behavior*, Third Edition (Princeton: Princeton University Press).

