Forthcoming in *Philosophy and Phenomenological Research*.

DECISION, CAUSALITY, AND PRE-DETERMINATION

Boris Kment
Princeton University

Abstract. Evidential decision theory (EDT) says that the choiceworthiness of an option depends on its evidential connections to possible outcomes. Causal decision theory (CDT) holds that it depends on your beliefs about its causal connections. While Newcomb cases support CDT, Arif Ahmed has described examples that support EDT. A new account is needed to get all cases right. I argue that an option A's choiceworthiness is determined by the probability that a good outcome ensues at possible A-worlds that match actuality in the facts causally unaffected by your decision (the "unaffected facts"). Moreover, you should evaluate A on the assumption that A is compossible with the unaffected facts. This view entails that you should use EDT when evaluating A on the assumption that the unaffected facts determine your action, but use CDT when assessing A on the opposite assumption. A's choiceworthiness equals a weighted average of these conditional assessments. The weights are determined by your beliefs about whether the unaffected fact determine your action. This account gets both Newcomb and Ahmed cases right. According to an influential view, whether you take the unaffected facts to determine your action can make a difference to whether you can regard yourself as free and the action as being under your control. While my account is neutral on this issue, it entails that whether you take the unaffected facts to determine your action is important in a different way: it matters to whether you should follow EDT or CDT.

Proponents of evidential decision theory (EDT) hold that you should rank your options by the strength of the evidence they provide for outcomes you value. Many decision theorists believe that this account yields the wrong result in so-called "Newcomb cases," in which the option that provides the best evidence for a good result does not causally promote that result. They conclude that making rational decisions requires attention not to evidential, but to causal relationships. Several elaborate accounts, collectively known as "causal decision theory" (CDT), have emerged from this thought. Their core thesis is often described, to a first approximation, as the idea that you should choose one of the options that you regard as most likely to cause valuable outcomes.

CDT faces its own set of apparent counterexamples, which have proliferated in the recent literature. I will focus on two such examples due to Arif Ahmed, which suggest that existing versions of CDT do poorly, and EDT does well, when applied to certain cases in which you are confident that your action is determined by the past and the natural laws. The goal of this paper is to diagnose the underlying error in orthodox CDT and to offer a new account that yields the right results in both Newcomb and Ahmed cases.

After describing EDT and Newcomb scenarios (§1), CDT (§2), and Ahmed cases (§3), I argue that existing versions of CDT mischaracterize the role of causal notions in rational choice (§4). What matters to the choiceworthiness of an option A is not the likelihood that A will cause a good outcome. Rather, what matters is the probability that a good outcome ensues at possible A-worlds where the facts that are causally unaffected by your decision are the way they actually are (§5). Moreover, you should evaluate A on the assumption that A is compossible with these unaffected facts (§6). In §7 I show that this account can be viewed as a hybrid between CDT and EDT, since it entails the following. When assessing your options on the assumption that your action is determined by the unaffected facts, you should follow EDT. When assessing them on the opposite assumption, you ought to follow CDT. Your overall evaluation should equal a weighted average of the two conditional assessments. The weights are determined by your beliefs about whether your action is determined by facts unaffected by your decision.

According to a much-discussed philosophical view, whether you take your actions to be determined by factors beyond your causal influence makes an important difference to whether you can view yourself as free. There has been much less discussion about whether it makes a difference to how you should make decisions. While my account is neutral on the former issue, it takes a stance on the latter: whether you take your actions to be determined by factors outside of your causal influence makes a difference to whether you should follow EDT or CDT.

I discuss and defend my theory's verdict about Ahmed cases in §6 and its predictions in Newcomb scenarios in §§8−9. My view and orthodox CDT agree on some types of Newcomb cases but disagree on others. Orthodox CDT's verdict is underwritten by a certain dominance principle. My account therefore requires us to reject this principle. I argue that we have independent reasons to do so, offer a diagnosis of the principle's failure, and formulate a restricted version of it that avoids the problem and is consistent with my theory (§8.2). §9 responds to an objection to my treatment of Newcomb cases. I conclude by comparing my account to other responses to Ahmed cases recently proposed by Alexander Sandgren and Timothy Williamson and by James Joyce (§10).

The Ahmed scenarios discussed in this paper are not the only apparent counterexamples to CDT. Other troublesome cases have been described by Egan (2007), Ahmed (2014a, 2021), Spencer and Wells (2019), and Spencer (2021a, 2021b), among others. However, I think that these other examples reveal problems for existing versions of CDT that differ from those brought to light by the cases discussed in this paper. I do not claim that the present version of my account can resolve these other difficulties. It will remain a task for future work to address them.

1. EDT and Newcomb

The most common approach in decision theory (which I will follow) represents your possible actions by pairwise metaphysically incompossible propositions called *options*; your creedal state by a subjective probability function Cr called *credence function*, whose sample space Ω_{Cr} is the set of worlds that are epistemically possible for you; and your values by a *value function* V from possible worlds to real numbers. The set of all possible worlds can be partitioned into equivalence classes of worlds, called *outcomes*, that have the same value. (For readability, I will suppress corner quotes

and quotation marks in quasi-quotation and quote names when there is no risk that doing so will lead to confusion. Variables A, B will range over options, w over worlds, O over outcomes. $V(O) = r$ will mean that $V(w) = r$ for all $w \in O$. For simplicity, I will pretend, when nothing hangs on it, that the set of possible worlds is finite.) We aim to define a *utility function*, U, that maps each option A to a real number that measures A's choiceworthiness. Most theories allow us to think of U(A) as a weighted average of the values of the different A-worlds. The weight of each A-world w equals the probability you assign to w under the supposition that you do A (Joyce 1999, Elga 2022). Let Cr^A be your probability function under the supposition A. Its sample space, Ω_{Cr^A}, contains only A-worlds. Letting $E(V, Cr^A)$ be the expectation of V relative to Cr^A, we can state the view thus:

(1) $U(A) = E(V, Cr^A) = \sum_O Cr^A(O)\, V(O)$

A is rationally permissible iff, for any option B, $U(A) \geq U(B)$.[1]

Cr^A is defined by its theoretical role, as given by formula (1). Decision theorists disagree about which probability function plays this role. According to EDT, Cr^A equals the probability distribution conditional on A, $Cr(- \mid A)$. Consequently, U(A) equals A's *evidential value*, EV(A), defined below.

EDT. $U(A) = EV(A) =_{def} \sum_O Cr(O \mid A)\, V(O)$

Roughly speaking, EDT tells you to choose the option that provides the best evidence for an outcome you value.

EDT confronts the following counterexample (Nozick 1969). (I will say that *you are certain of P* iff P holds at every world in Ω_{Cr}.)

> *Newcomb*. During a game show, a transparent box containing \$1,000 (\$K) and an opaque box are placed in front of you. You are certain that you will receive the opaque box as a gift and that either s_0 or s_M holds.
>
> s_M: The opaque box contains \$1,000,000 (\$M).
> s_0: The opaque box contains \$0.
>
> You have two options.
>
> One-boxing (B_1): You do not take the transparent box.
> Two-boxing (B_2): You take the transparent box.
>
> You know the following with certainty:
>
> Yesterday, a very reliable oracle predicted your action. A deterministic mechanism then ensured that the opaque box would contain \$1,000,000 if the oracle predicted B_1, and \$0 if she predicted B_2. There is no backwards causation, so your decision does not causally influence the prediction.

[1] Or at least, this is true if the number of options is finite, as is the case in all examples I will discuss.

Moreover, you are 99% confident, both unconditionally and conditionally on either option, that the oracle's prediction was correct. I will assume (throughout this paper) that V(You receive $\$x$) = x.

Since $Cr(s_M \mid B_1) = Cr(s_0 \mid B_2) = .99$, B_1 is excellent evidence for the claim that you will receive \$1,000,000 and B_2 is excellent evidence against it. EDT therefore entails that B_1 is uniquely rational:

$$EV(B_1) = Cr(\$M \mid B_1)\, V(\$M) + Cr(\$0 \mid B_1)\, V(\$0) = .99 \times 1{,}000{,}000 + .01 \times 0 = 990{,}000$$
$$EV(B_2) = Cr(\$(M+K) \mid B_2)\, V(\$(M+K)) + Cr(\$K \mid B_2)\, V(\$K)$$
$$= .01 \times 1{,}001{,}000 + .99 \times 1{,}000 = 11{,}000$$

The following argument (spelled out more fully in §8) convinced many philosophers that this is wrong.

Dominance Argument. Your action makes no difference to the content of the opaque box, and no matter what it contains, the outcome will be better if you two-box than if you one-box. This is usually described by saying that two-boxing *dominates* one-boxing. If one option dominates another, you should prefer the former option to the latter. You therefore ought to two-box.

Some find the following related argument very compelling.

Better-Informed Self Argument (cp. Nozick 1969: 116–17). Suppose you knew what was in the opaque box. Then you would either know that it contains \$1,000,000, or know that it contains nothing. Either way, you would know that two-boxing would yield a greater payoff than one-boxing. Therefore, without knowing what is in the opaque box, you can be sure that you would two-box if you had that knowledge. Moreover, it seems plausible that in the actual situation you should pick the option that you know you would choose if you had information about the content of the opaque box—you should defer to your hypothetical better-informed self. Hence, you ought to two-box.

2. CDT

According to EDT, $Cr^A(O)$ equals $Cr(O \mid A)$ and therefore reflects a doxastic or evidential connection between O and A. CDTists take Newcomb to show that this view can yield the wrong result when an option provides good evidence for a result that it does not causally promote. (One-boxing provides evidence that you will become a millionaire without causally contributing to this result.) Their diagnosis is that EDT pays insufficient attention to the role of causal beliefs in rational choice. CDTists typically characterize this role as follows:

Optimal Effect. You should choose one of the options whose causal effects have the highest expected value.[2]

In Newcomb, you are certain that neither option causally promotes getting \$1,000,000 and neither option causally promotes not getting \$1,000,000, but that B_2 causally promotes getting the \$1,000 while B_1 causally promotes not getting that money. You therefore expect B_2's causal effects to be more beneficial than B_1's. According to Optimal Effect, you ought to two-box.

By CDT's lights, the probability function Cr^A in formula (1) must reflect your beliefs about how strongly the different options causally promote the various outcomes. The best known form of CDT, *counterfactual decision theory*, defines Cr^A in terms of counterfactual conditionals or subjunctive probabilities (Stalnaker 1978, Gibbard and Harper 1978, Lewis 1981, Joyce 1999).[3] On the simplest such account (Gibbard and Harper 1978), you should choose one of the options that maximize what we may call *expected counterfactual value* (defined below).

A's *counterfactual value*, *CV(A)*, is the value of the outcome that would result if you were to do A. A's *expected counterfactual value*, *ECV(A)*, is your expectation of CV:

$$ECV(A) =_{def} E(CV(A), Cr) = \sum_O Cr(A \mathrel{\square\!\!\rightarrow} O)\, V(O)$$

ECV-Maximization (ECV-Max). $U(A) = ECV(A)$

According to ECV-Max:

(2) $Cr^A(O) = Cr(A \mathrel{\square\!\!\rightarrow} O)$

Counterfactual decision theorists usually understand counterfactuals in accordance with the standard possible-worlds account (Stalnaker 1968, Lewis 1973). Roughly speaking, $A \mathrel{\square\!\!\rightarrow} O$ is true at possible world *w* iff O holds at all the A-worlds closest to *w*. CV(A) exists only if $A \mathrel{\square\!\!\rightarrow} O$ is true for some option O. ECV(A) is defined only if you are certain that CV(A) exists. ECV-Max therefore assumes (3).

(3) For every option A, you are certain that there is a unique outcome that obtains at all the closest A-worlds.

To derive predictions from ECV-Max, we need to know what the A-worlds closest to a given world *w* look like. Counterfactual decision theorists commonly assume that, if A is a proposition about time *t*, then the A-worlds closest to *w* by and large meet the following conditions:

(4) (i) they are like *w* before *t* (except perhaps in cases of backwards causation in which events before *t* are actually causally influenced by whether A holds), and
(ii) they conform to *w*'s laws.

[2] For statements of CDT along these lines (by friends and foes of CDT), see e.g. Lewis 1981: 5, Egan 2007: 94, Ahmed 2015: 262.
[3] For formulations of CDT without counterfactuals, see Skyrms 1980: §IIC and Edgington 2011.

However, if w is deterministic and ¬A holds at w, then any initial segment of w's history, combined with w's laws, necessitates ¬A. In that case, no possible A-world meets both of the criteria (4)(i)–(ii) perfectly. The A-worlds closest to w are then the ones that provide the best trade-off between these criteria. Lewis 1979 gives a detailed account of this trade-off: barring backward causation, the closest A-worlds are like w until t, or until shortly before t. They then diverge smoothly from w so as to make A true. If w is deterministic, then that requires a small violation of w's laws (a small "miracle"). After the miracle, the A-worlds closest to w conform perfectly to w's laws. This is called a *non-backtracking account*, since it entails that the pre-t history is almost completely counterfactually independent of A. Unless otherwise noted, I will assume that this account is correct.

Combined with a non-backtracking account, ECV-Max yields the desired result that you should two-box in Newcomb. Since you are certain that the content of the opaque box was decided yesterday and that there is no backwards causation, you are certain that s_M and s_0 are counterfactually independent of the options. So, you are certain that, if s_M is true, then so are $B_1 \mathbin{\square\!\!\rightarrow} s_M$ and $B_2 \mathbin{\square\!\!\rightarrow} s_M$; and if s_0 is true, then so are $B_1 \mathbin{\square\!\!\rightarrow} s_0$ and $B_2 \mathbin{\square\!\!\rightarrow} s_0$. So, $Cr(B_1 \mathbin{\square\!\!\rightarrow} \$M) = Cr(B_2 \mathbin{\square\!\!\rightarrow} \$(M+K)) = Cr(s_M)$ and $Cr(B_1 \mathbin{\square\!\!\rightarrow} \$0) = Cr(B_2 \mathbin{\square\!\!\rightarrow} \$K) = Cr(s_0) = 1 - Cr(s_M)$. By ECV-Max:

$$\begin{aligned} U(B_1) &= Cr(B_1 \mathbin{\square\!\!\rightarrow} \$M)\ 1{,}000{,}000 + Cr(B_1 \mathbin{\square\!\!\rightarrow} \$0)\ 0 \\ &= Cr(s_M)\ 1{,}000{,}000 \end{aligned}$$

$$\begin{aligned} U(B_2) &= Cr(B_2 \mathbin{\square\!\!\rightarrow} \$(M+K))\ 1{,}001{,}000 + Cr(B_2 \mathbin{\square\!\!\rightarrow} \$K)\ 1{,}000 \\ &= Cr(s_M)\ 1{,}001{,}000 + (1 - Cr(s_M))\ 1{,}000 \\ &= Cr(s_M)\ 1{,}000{,}000 + 1{,}000 \end{aligned}$$

Hence, $U(B_2) > U(B_1)$.

As Gibbard and Harper note, ECV-Max has a significant limitation: (3) can fail in a number of ways. For example, you might believe that the outcome is partly determined by chance processes that turn out differently across the closest A-worlds, so that $A \mathbin{\square\!\!\rightarrow} O$ is false for every outcome O. (In such cases, you think that there are different outcomes that *might* result if you did A, but none that *would* come about.) Lewis (1981) and James Joyce (1999) offer versions of CDT that do not rest on (3) but that agree with ECV-Max when (3) holds. Since (3) is true in all counterexamples to ECV-Max considered in this paper, they are also counterexamples to Lewis's and Joyce's theories. We can therefore simplify the discussion by focusing wholly on ECV-Max.

3. Counterexamples to CDT

Consider two counterexamples to CDT:

Past-Bet$_1$ (Ahmed 2013a, 2014b: §5.1). You are deciding between raising your arm (Arm) and not raising it (¬Arm). Each option amounts to accepting a bet on proposition P_1, with the payoffs displayed in Table 1.

	P_1	$\neg P_1$
Arm	\$10	−\$1
¬Arm	\$1	−\$10

Table 1. Past-Bet$_1$.

P_1: The state of the universe yesterday at noon, combined with the natural laws of the actual world, metaphysically necessitates ¬Arm.

(In P_1, "the actual world" rigidly refers to the world that is in fact actualized.) You are certain that determinism holds and that you have no causal influence on the past or on what the actual laws are, and therefore have no causal influence on P_1's truth-value.[4]

Since you are certain of determinism, you are certain of the following: P_1 holds iff ¬Arm is true.

Ahmed claims that ¬Arm is uniquely rational, and I agree. Choosing Arm seems relevantly like taking a bet on the claim that you are not taking that very bet, while ¬Arm seems relevantly like taking a bet on the claim that you are taking that bet. The first action is self-undermining, the second highly recommendable. However, ECV-Max (combined with the non-backtracking account of counterfactuals) entails that you should choose Arm. Since P_1 is about yesterday, $Cr(Arm \mathrel{\square\!\!\rightarrow} P_1) = Cr(\neg Arm \mathrel{\square\!\!\rightarrow} P_1) = Cr(P_1)$ and $Cr(Arm \mathrel{\square\!\!\rightarrow} \neg P_1) = Cr(\neg Arm \mathrel{\square\!\!\rightarrow} \neg P_1) = Cr(\neg P_1) = 1 - Cr(P_1)$. So:

$$\begin{aligned}
ECV(Arm) &= Cr(Arm \mathrel{\square\!\!\rightarrow} \$10)\ 10 + Cr(Arm \mathrel{\square\!\!\rightarrow} -\$1)\ (-1) \\
&= Cr(Arm \mathrel{\square\!\!\rightarrow} P_1)\ 10 + Cr(Arm \mathrel{\square\!\!\rightarrow} \neg P_1)\ (-1) \\
&= Cr(P_1)\ 10 + (1 - Cr(P_1))\ (-1) \\
&= 11\ Cr(P_1) - 1 \\
ECV(\neg Arm) &= Cr(\neg Arm \mathrel{\square\!\!\rightarrow} \$1)\ 1 + Cr(\neg Arm \mathrel{\square\!\!\rightarrow} -\$10)\ (-10) \\
&= Cr(P_1) + (1 - Cr(P_1))\ (-10) \\
&= 11\ Cr(P_1) - 10
\end{aligned}$$

Law-Bet$_1$ (Ahmed 2013b, 2014b: §5.2). You are highly confident that a certain principle L is true, and certain that, if L is true, then L is a natural law. Moreover, L is deterministic in the following sense:

(5) If two possible L-worlds are in the same state at any time, then they have identical histories.

Your options are to assert L ($Assert_L$) and to assert ¬L ($Assert_{\neg L}$). You assign value 1 to asserting a truth (True) and value 0 to asserting a falsehood (False). Nothing else is at stake. The options are probabilistically independent of L in Cr, and you are certain that you have no causal influence on L's truth-value.

Ahmed claims that you should choose $Assert_L$ since you are confident that L is true. Although I will argue in §6.2 that that is not true of *all* versions of Law-Bet$_1$, I agree that it is true in some. However, ECV-Max, when combined with the non-backtracking account, wrongly predicts that $U(Assert_{\neg L}) \geq U(Assert_L)$ in all versions of Law-Bet$_1$. To begin with:

[4] Ahmed merely stipulates that you are highly confident, not that you are certain, of determinism. However, focusing on a version of the example in which you are certain of determinism will simplify the discussion. Nothing important hinges on this decision.

$ECV(Assert_L) = Cr(Assert_L \boxrightarrow True)\ 1 + Cr(Assert_L \boxrightarrow False)\ 0 = Cr(Assert_L \boxrightarrow True)$
$ECV(Assert_{\neg L}) = Cr(Assert_{\neg L} \boxrightarrow True)\ 1 + Cr(Assert_{\neg L} \boxrightarrow False)\ 0 = Cr(Assert_{\neg L} \boxrightarrow True)$

Moreover, $Assert_L \boxrightarrow True$ entails $Assert_{\neg L} \boxrightarrow True$. (*Proof*. Suppose $Assert_L \boxrightarrow True$ holds. Then L is true at the closest $Assert_L$-worlds. Let proposition S completely describe the actual state of the universe yesterday at noon. By the non-backtracking account, S also holds at the closest $Assert_L$-worlds. By (5), all possible S&L-worlds have the same history as the closest $Assert_L$-worlds and therefore make $Assert_L$ true. By the non-backtracking account, S holds at the closest $Assert_{\neg L}$-worlds. Since these worlds do *not* make $Assert_L$ true, ¬L must hold at them. Hence, $Assert_{\neg L} \boxrightarrow True$ is true.) Thus, $U(Assert_L) = Cr(Assert_L \boxrightarrow True) \leq Cr(Assert_{\neg L} \boxrightarrow True) = U(Assert_{\neg L})$.

In deriving the claim that $U(Assert_{\neg L}) \geq U(Assert_L)$, I appealed to the non-backtraking account in addition to ECV-Max. Defenders of ECV-Max might claim that the problem lies not with ECV-Max, but with the non-backtracking view. We could reject this view in favor of (6) (Goodman 2015, Dorr 2016).[5]

(6) The A-worlds closest to *w* are possible worlds that have the same laws as *w* and conform to these laws perfectly.

According to (6), if determinism and ¬A are true, then the closest A-worlds differ from actuality throughout their pre-antecedent histories. However, these pre-antecedent differences might be very small. ECV-Max-cum-(6) gets Past-Bet_1 and Law-Bet_1 right.

Nevertheless, adopting (6) does not help ECV-Maxists, since ECV-Max-cum-(6) confronts its own counterexamples, including a case described by Williamson and Sandgren (2021: §5.1.1) and the following example.

Law-Bet_2. Both Arm and ¬Arm amount to accepting a bet on proposition P_2, with the payoffs displayed in Table 2.

P_2: The natural laws, combined with the state of the universe yesterday at noon at the actual world, metaphysically necessitates ¬Arm.

(In P_2, "the actual world" is rigid.) You are certain that determinism holds and that you have no causal influence on the past or on what the actual laws are.

	P_2	$\neg P_2$
Arm	\$10	−\$1
¬Arm	\$1	−\$10

Table 2. Law-Bet_2.

You should choose ¬Arm, for the same reason as in Past-Bet_1. But ECV-Max-cum-(6) entails the opposite. (6) entails that, if P_2 holds, then $Arm \boxrightarrow P_2$ and $\neg Arm \boxrightarrow P_2$ are true, and if $\neg P_2$ holds, then $Arm \boxrightarrow \neg P_2$ and $\neg Arm \boxrightarrow \neg P_2$ are true. Hence, $Cr(Arm \boxrightarrow P_2) = Cr(\neg Arm \boxrightarrow P_2) = Cr(P_2)$ and $Cr(Arm \boxrightarrow \neg P_2) = Cr(\neg Arm \boxrightarrow \neg P_2) = Cr(\neg P_2)$. According to ECV-Max-cum-(6), $U(Arm) = 10\ Cr(Arm \boxrightarrow P_2) - Cr(Arm \boxrightarrow \neg P_2) = 10\ Cr(P_2) - Cr(\neg P_2) > Cr(P_2) - 10\ Cr(\neg P_2) = Cr(\neg Arm \boxrightarrow P_2) - 10\ Cr(\neg Arm \boxrightarrow \neg P_2) = U(\neg Arm)$.

[5] Cp. Nute 1980, Bennett 1984, Loewer 2007, Albert 2015, Wilson 2014.

4. Two Views on Causation and Rational Choice

Existing versions of CDT enshrine Optimal Effect. ECV-Maxists add the further assumption that your decision-relevant causal beliefs amount to credences in certain counterfactuals: your estimate of A's efficacy in causing O equals Cr(A $\square\!\!\rightarrow$ O). Law-Bet$_1$ seems to cast doubt on the latter assumption. Given your confidence in L's truth, you should take Assert$_L$, but not Assert$_{\neg L}$, to be highly efficacious in bringing about True, and yet Cr(Assert$_L$ $\square\!\!\rightarrow$ True) $\leq$ Cr(Assert$_{\neg L}$ $\square\!\!\rightarrow$ True). Proponents of Optimal Effect might say that it is due to this mismatch between causal and counterfactual beliefs, rather than to the falsity of Optimal Effect, that ECV-Max gets the example wrong. That there should be such a mismatch is hardly surprising, given the well-known difficulties that confront attempts to state necessary and sufficient conditions for causation in counterfactual terms (see Collins, Hall and Paul 2004).

Once we accept that causal and counterfactual beliefs can come apart, we could try to accommodate Optimal Effect by formulating CDT directly in terms of your beliefs about causal efficacies, without appeal to counterfactuals. (Admittedly, this move by itself does not obviously solve the problem that Past-Bet$_1$ presents for CDT. But one might hope that it addresses at least *some* Ahmed counterexamples and can therefore be *part* of the solution.) However, there are examples that show that such an account would yield worse results than counterfactual decision theory in some cases. Consider the following example (which has the same structure as a case described in Hitchcock 2013).[6]

> *Button*. You have a button on your desk and so does Mary. Your options are to press your button before 8 p.m. and not to press it. (Your button will disappear at 8 p.m.) You are certain of the following: if either button is pressed today, the Department of World Improvement will see to it that the world will improve the next morning. Once the Department has been notified that one of the buttons has been pushed, the buttons are disconnected and later button pushes are no longer registered. All this will happen by some deterministic mechanism over which you have no influence. If neither you nor Mary presses the button, the world will not improve. You value world improvement but do not care about what causes it. Nothing else you care about is at stake. You are certain that Mary will press her button shortly after 8 p.m. (her button will not disappear until midnight) and that there is no causal connection between your decision and her action.

If you push your button, your action will be a preempting cause of the ensuing world improvement (it will preempt a second potential cause, namely Mary's button pushing). As in other preemption cases, causation and counterfactual dependence come apart: you are certain that whether the world improves is counterfactually independent of your action. And yet, you are also certain that if you push your button, your action will cause world improvement, and if you do not push, your action (not pushing) will not cause world improvement. Optimal Effect, understood in genuinely causal

[6] Hitchock believes that there is a counterfactual notion of causation (called "causal dependence") as well as another concept ("actual causation") that resist counterfactual analysis. (Cp. Hall 2004.) What I describe as the causal and counterfactual interpretations of Optimal Effect would be described by him as an interpretation in terms of actual causation and an interpretation in terms of causal dependence, respectively.

rather than counterfactual terms, predicts that pushing is uniquely rational, since pushing is more likely than not pushing to cause the world to improve. That prediction seems wrong, however. Since you are certain that the world will improve no matter what you do, not pressing is as good as pressing. (By hypothesis, you do not care about what causes the world to improve.) Note that ECV-Max gets the case right: ECV(Push) = ECV(¬Push), since Cr(Push □→ World Improvement) = Cr(¬Push □→ World Improvement) = 1.

Proponents of Optimal Effect confront a dilemma. If they identify agents' decision-relevant causal beliefs with their credences in counterfactuals, they face counterexamples like Past-Bet_1 and Law-Bet_1. If they instead distinguish between causal and counterfactual judgments and formulate their decision theory in terms of non-counterfactual causal beliefs, they get Button wrong. It is therefore far from clear that a viable version of Optimal Effect exists. There is good motivation to look for an alternative view.

I propose that the significance of causal notions in rational choice is not as described by Optimal Effect. Instead, it can be captured, to a first approximation, by the following highly plausible principle. When deliberating about what to do, you need to accept (hold fixed) the facts that you take to be causally unaffected by your decision. You should ask, "Given these facts, what is my best option?". More precisely, on the assumption that proposition P is true and that P describes matters that are causally unaffected by your decision, ¬P-worlds are irrelevant to a rational assessment of your options. Reflection on such worlds is a form of wishful thinking that has no place in rational choice.

Let me state these ideas more precisely. First, some definitions. I will use "fact" for true propositions. Where $A_1, \ldots, A_n$ are your options and A_i is the option that you in fact choose, I will say that P *is causally downstream from your decision* iff (a) P is true, and (b) either the fact that you form the intention to do A_i stands in the ancestral relation of causation to P, or for some $j \neq i$, the fact that you do not form the intention to do A_j stands in the ancestral relation of causation to P.

> A fact P is *(causally) unaffected (by your decision)* iff_{def} (i) P is not causally downstream from your decision, and (ii) both P and ¬P are distinct from each of your options.[7]
>
> A proposition P is *unaffected* iff_{def} either P or ¬P is an unaffected fact.

The unaffected facts include not only certain matters of particular fact, but also the natural laws (since the laws are not causally downstream from your decision). My account of the role of causal notions in decision-making can be stated as follows.

> *Fixity*. For any option A, if you are certain that P is an unaffected fact, then on the supposition that A holds, you are certain of P (i.e., P holds at all worlds in Ω_{Cr_A}).

[7] The notion of distinctness in clause (i) might have to be stronger than mere non-identity. In particular, it might require that, for any option A, neither A nor ¬A is partially (metaphysically) grounded in P or in ¬P, and neither P nor ¬P is partially grounded in A or in ¬A. (For an introduction to grounding, see Schaffer 2009, Rosen 2010, Koslicki 2012, Audi 2012, Fine 2012.) By this criterion, A and P are not distinct if, say, A is *I raise both arms* and P is *I raise my left arm*.

Fixity entails Guaranteed Outcome.[8]

> *Guaranteed Outcome*. If (i) you are certain that P is an unaffected fact and (ii) A & P metaphysically necessitates a unique outcome O, then U(A) = V(O).

Two clarifications about the notion of unaffectedness underlying my account are called for. *First*, I take it to be very plausible that there is a notion of causal unaffectedness that makes the following principle true.

> *Past−Law Unaffectedness*. Necessarily, if there are no instances of backwards causation, and if proposition P is solely about the time before your decision and/or about the laws of nature, then P is causally unaffected by your decision.

My account should be understood as involving a notion of causal unaffectedness that makes this principle true. Such a notion is hyperintensional, i.e. it applies to propositions (such as Russellian structured propositions) that are individuated more finely than sets of possible worlds. To see this, note that, according to Past–Law Unaffectedness, there could be pairs of necessarily equivalent true propositions one of which is an unaffected fact while the other is not. For illustration, suppose that determinism holds, i.e. that the laws and the state of the universe at any given time metaphysically necessitate all truths about history. Assume further that there is no backwards causation. Let D be the conjunction of the actual laws and let H_y (H_t) completely describe the state of the universe at some moment yesterday (tomorrow). D & H_y and D & H_t are necessarily equivalent. By Past–Law Unaffectedness, D & H_y is causally unaffected by your decision. However, D & H_t is partly about tomorrow. It may therefore describe some matters that are causally downstream of your decision, so that D & H_t is not an unaffected fact. *Secondly*, the concept of causation underlying the notion of unaffectedness should be understood broadly, so as to include relationships of grounding and metaphysical explanation.

I will henceforth use "state" for sets of possible worlds while reserving "proposition" for the more fine-grained entities. Since options and decisions are among the relata of causal affectedness, I will take them to be propositions (of the forms *I do X* and *I decide to do X*, respectively, where X is a type of action) rather than states. I will continue to treat outcomes as states. Given any proposition P, "⟨P⟩" will stand for the set of all possible worlds at which P holds. For readability, I will write "Cr(P)" for Cr(⟨P⟩), "probabilistic dependence on P" for probabilistic dependence on ⟨P⟩, etc. Given a set X each of whose members is either a state or proposition, let X* be the set obtained by replacing every proposition Q in X with ⟨Q⟩. (X* contains only states.) I will say that X *metaphysically necessitates* state *s* iff $(\cap X^*) \subseteq s$, and that X *metaphysically necessitates* proposition P iff X metaphysically necessitates ⟨P⟩. X's members are *(mutually) compossible* iff $(\cap X^*) \neq \emptyset$. Given a state *s* and proposition P, I will write *s* & P ($s \vee P$) for $s \cap \langle P \rangle$ ($s \cup \langle P \rangle$).

[8] Suppose Fixity and clauses (i)−(ii) of Guaranteed Outcome hold. By Fixity and (i), all worlds in Ω_{Cr^A} are P-worlds. Since all worlds in Ω_{Cr^A} are A-worlds, it follows by (ii) that all worlds in Ω_{Cr^A} are O-worlds. By (1), U(A) = V(O).

Button and Past-Bet$_1$ illustrate Fixity and Guaranteed Outcome. *Button*. You are certain that (7) is an unaffected fact, and that (7) and certain other unaffected facts (including both the laws and certain matters of particular fact) together metaphysically necessitate world improvement.

(7) Mary will push her button before midnight.

By Guaranteed Outcome, U(Push) = U(¬Push) = V(World improvement). *Past-Bet$_1$*. You are certain that (8) is an unaffected fact. Hence, by Fixity, (9) holds.

(8) ¬Arm nomically necessitates P_1, and Arm nomically necessitates $\neg P_1$.

(9) $Cr^{Arm}((8)) = Cr^{\neg Arm}((8)) = 1$

Therefore, $Cr^{Arm}(\neg P_1) = Cr^{\neg Arm}(P_1) = 1$, and $U(\neg Arm) = 1 > -1 = U(Arm)$.

ECV-Max's problematic verdicts about Past-Bet$_1$ and Law-Bet$_1$ are due to the fact that it violates Fixity. *Past-Bet$_1$*. ECV-Max entails that $Cr^{Arm}((8)) \leq Cr^{Arm}(\neg P_1) = Cr^{Arm}(-\$1) = Cr(Arm \boxrightarrow -\$1) = Cr(\neg P_1)$ and $Cr^{\neg Arm}((8)) \leq Cr^{\neg Arm}(P_1) = Cr^{\neg Arm}(\$1) = Cr(\neg Arm \boxrightarrow \$1) = Cr(P_1)$, which entails that either $Cr^{Arm}((8)) < 1$ or $Cr^{\neg Arm}((8)) < 1$. That contradicts (9) and therefore violates Fixity. This violation accounts for ECV-Max's false prediction about Past-Bet$_1$. *Law-Bet$_1$*. Let $U(- \mid L)$ ($U(- \mid \neg L)$) be the utility function of a possible agent whose credence function is $Cr(- \mid L)$ ($Cr(- \mid \neg L)$) and who has the same two options, and assigns the same values to True and False, as you. $Assert_L$ and L (¬L) necessitates True (False) while $Assert_{\neg L}$ and L (¬L) necessitates False (True). (10) follows by Guaranteed Outcome (and hence by Fixity).

(10) $U(Assert_L \mid L) = 1$ $\quad U(Assert_{\neg L} \mid L) = 0$
$\quad U(Assert_L \mid \neg L) = 0$ $\quad U(Assert_{\neg L} \mid \neg L) = 1$

However, ECV-Max can be shown to entail:

(11) (a) $U(Assert_L \mid L) = Cr(Assert_L \boxrightarrow True \mid L) \leq Cr(Assert_L \mid L)$
(b) $U(Assert_{\neg L} \mid L) = Cr(Assert_{\neg L} \boxrightarrow True \mid L) \geq Cr(Assert_L \mid L)$ [9]

By (11), $U(Assert_{\neg L} \mid L) \geq U(Assert_L \mid L)$. Not only is this conclusion absurd (conditional on L, you should surely assign a higher utility to asserting L than to asserting ¬L!), but it also contradicts (10) and therefore violates Fixity. This violation accounts for ECV-Max's mistaken prediction about Law-Bet$_1$. For, ECV-Max entails:

(12) (a) $U(Assert_L) = Cr(L)\, U(Assert_L \mid L) + Cr(\neg L)\, U(Assert_L \mid \neg L)$

[9] *Proof that ECV-Max entails (11)(a)*. Suppose $L \,\&\, Assert_{\neg L}$ holds at *w*. Since the $Assert_L$-worlds closest to *w* are like *w* yesterday, it follows by (5) that they violate L. Hence, $\neg(Assert_L \boxrightarrow True)$ holds at *w*. This shows that $L \,\&\, Assert_{\neg L}$ necessitates $\neg(Assert_L \boxrightarrow True)$. Hence, $L \,\&\, (Assert_L \boxrightarrow True)$ necessitates $\neg Assert_{\neg L}$. Therefore, $Cr(Assert_L \boxrightarrow True \mid L) \leq Cr(\neg Assert_{\neg L} \mid L) = Cr(Assert_L \mid L)$. Given ECV-Max, (11)(a) follows. *Proof that ECV-Max entails (11)(b)*. By (5) and the non-backtracking account, $Assert_L \,\&\, L$ entails $Assert_{\neg L} \boxrightarrow \neg L$ and therefore entails $Assert_{\neg L} \boxrightarrow True$.

(b) $U(\text{Assert}_{\neg L}) = Cr(L)\ U(\text{Assert}_{\neg L} \mid L) + Cr(\neg L)\ U(\text{Assert}_{\neg L} \mid \neg L)$ [10]

Combined with (10), (12) yields the result that $U(\text{Assert}_L) = Cr(L) \geq Cr(\neg L) = U(\text{Assert}_{\neg L})$. Hence, it is only because ECV-Max violates Fixity by rejecting (10) that it wrongly predicts that $U(\text{Assert}_{\neg L}) \geq U(\text{Assert}_L)$.

I will use Fixity to formulate a new decision theory (§§5−7) and argue that it yields the right results in both Ahmed cases (§6.2) and Newcomb examples (§§8−9).

5. Real value and rational choice

In formulating my view, I will restrict my attention to cases in which you are certain of Pre-Determined Outcome.

> *Predetermined Outcome*. For every option A, A and the unaffected facts together metaphysically necessitate an outcome.

(Predetermined Outcome holds in all examples considered in this paper.) I think it would be relatively straightforward to generalize my account to cases in which you are not certain of Predetermined Outcome. For brevity's sake, I will leave this task for another occasion.

I will start by formulating an account that is restricted to scenarios in which you are certain that the proposition ♦A holds for every option A.

(♦A) The unaffected facts (including both the laws and the unaffected matters of particular fact) are jointly metaphysically compossible with A.

If there is no backwards causation, then all facts about the history before your decision are unaffected facts. If that is so and if there is more than one option, then ♦A can be true for every option A only if indeterminism holds. In §6, I will generalize my account to cases in which ♦A does not hold for all options, including deterministic scenarios. (Note that I will leave it open whether the truth of ♦A is required for you to be free to do A, i.e. I will remain neutral in the dispute between compatibilists and incompatibilists.)

In motivating my account, I will use both Fixity and the following assumption (which we have already seen at work in the Better-Informed Self Argument of §1): when evaluating your options, it is rational to defer to a possible version of yourself who has more relevant information than you. More information about what? The unaffected facts. How much more information? There is no reason to impose any restriction. After all, if you have to choose between deferring to someone with complete information of the unaffected facts and deferring to someone with incomplete such information, you cannot go wrong by deferring to the fully-informed person. More precisely:

---

[10] The proof of (12)(a) from ECV-Max is below. The proof of (12)(b) from ECV-Max is analogous.

$$\begin{aligned} U(\text{Assert}_L) &= ECV(\text{Assert}_L) && \text{ECV-Max} \\ &= Cr(L)\ ECV(\text{Assert}_L \mid L) + Cr(\neg L)\ ECV(\text{Assert}_L \mid \neg L) && \text{Law of total expectation} \\ &= Cr(L)\ U(\text{Assert}_L \mid L) + Cr(\neg L)\ U(\text{Assert}_L \mid \neg L) && \text{ECV-Max} \end{aligned}$$

You should aim to assess your options in the same way as a possible agent, You*, who satisfies conditions (i)–(iii).

(i) You* is perfectly rational.
(ii) You* has the same options and value function as you.
(iii) You*'s credence function comes from your credence function by conditioning on true and complete information about the unaffected facts.

Being uncertain about what the unaffected facts are, you are also uncertain about what You*'s credence function is, and so you cannot know for sure what utility You* assigns to each option (call that utility U*(A)). The rule "assign the same utilities as You*" therefore cannot guide you as it stands. What you *can* do, however, is to assign to each option A your best estimate of U*(A). Your best estimate is your expectation of U*(A), i.e. E(U*(A), Cr). This suggests the following.

U-Max*. $U(A) = E(U^*(A), Cr)$

From U*-Max and Fixity we can derive a formula for computing utilities. Call state *s epistemically possible* iff $s \cap \Omega_{Cr} \neq \emptyset$, i.e. iff you are not certain of $\neg s$. A proposition P is *epistemically possible* iff $\langle P \rangle$ is epistemically possible. Partition the set of all possible worlds into classes of worlds at which the unaffected facts are the same. Call the cells of this partition *background states*. (Background states are maximally specific possibilities concerning what the unaffected facts are like.) Let $\mathcal{B}$ be the set of all epistemically possible background states. Let $b \in \mathcal{B}$ and let A be an option. Since you are certain of ♦A, b must be metaphysically compossible with A. Given that you are also certain of Predetermined Outcome, some outcome O_{Ab} must be metaphysically necessitated by $A \& b$. If b is true, then You*'s credence function is $Cr(- \mid b)$. You* is then certain that A and the unaffected facts together necessitate O_{Ab}, from which it follows by Guaranteed Outcome (and thus by Fixity) that $U^*(A) = V(O_{Ab})$. So, $E(U^*(A), Cr(- \mid b)) = V(O_{Ab})$. (13) follows by the law of total expectation. U*-Max and (13) entail (14).

(13) $E(U^*(A), Cr) = \sum_{b \in \mathcal{B}} Cr(b)\, E(U^*(A), Cr(- \mid b)) = \sum_{b \in \mathcal{B}} Cr(b)\, V(O_{Ab})$

(14) $U(A) = \sum_{b \in \mathcal{B}} Cr(b)\, V(O_{Ab})$

An alternative formulation of this account will be useful. Roughly speaking, my proposal is that you should choose the option from which you can expect the best outcome, given what the unaffected facts are like. To make this more precise, let us define a new quantity.

A's *real value*, *RV(A)*, is the value of the outcome that obtains at all possible A-worlds that match actuality perfectly in the unaffected facts.

(In this definition, "actuality" should be understood *non*-rigidly: at every world w, it picks out w.) RV(A) exists at possible world w iff (i) there are possible A-worlds that match w perfectly in the unaffected facts and (ii) the same outcome obtains at all these worlds. ♦A entails (i), and Predetermined Outcome entails (ii). Since you are certain of ♦A and of Predetermined Outcome, you are

certain that RV(A) exists. Let A's *Expected Real Value*, *ERV(A)*, be your expectation of RV(A), i.e. $\mathrm{ERV}(A) =_{\mathrm{def}} E(\mathrm{RV}(A), \mathrm{Cr})$. You should maximize ERV.

ERV-Max. U(A) = ERV(A)

ERV-Max is equivalent to U*-Max. For, given ♦A and Pre-Determined Outcome, Fixity entails that RV(A) = U*(A). (Let b^* be a true background state. By ♦A, there are A-worlds in b^*. These worlds are all and only the possible A-worlds that match actuality perfectly in the unaffected facts. By Pre-Determined Outcome, the same outcome $O_{A b^*}$ obtains at all of these worlds. So, $\mathrm{RV}(A) = V(O_{A b^*})$. Moreover, You*'s credence function is $\mathrm{Cr}(-\,|\,b^*)$. By Guaranteed Outcome, $U^*(A) = V(O_{A b^*})$.) Since you are certain of ♦A and Pre-Determined Outcome, you are certain that RV(A) = U*(A). Therefore:

(15) $\mathrm{ERV}(A) = E(U^*(A), \mathrm{Cr}) = \sum_{b \in \mathcal{B}} \mathrm{Cr}(b)\, V(O_{A b})$

Some philosophers distinguish between objective and subjective senses of "ought" (e.g., Jackson 1991, also see Railton 1986). While this distinction can be drawn for *oughts* of different flavors, I will focus on those of instrumental rationality. Suppose you know that a fair six-sided die was cast; if it landed on one, box X contains $1,000 and box Y is empty; otherwise, Y contains $1,000 and X is empty. Unbeknownst to you, the die landed on one. You have to choose between taking X and taking Y. You *objectively* ought to take X, since X contains the money. But you *subjectively* ought to take Y, since Cr(Y contains $1,000) = ⅚. (Decision theory studies this subjective *ought*.)

ECV-Maxists and ERV-Maxists can account for the two oughts by using a notion of the *objective (instrumental) value* of an option (Ahmed and Spencer 2020). Objective value needs to satisfy two conditions.

(16) The objective value of an option A measures the degree to which A tends, in the relevant circumstances, to promote what the agent regards as valuable.

(17) Facts about what it is rational to do in a given decision situation can be explained by the fact that rational agents aim to maximize objective value.

There are different possible measures of an option's tendency to promote value in the relevant circumstances that differ in the conditions under which circumstances count as relevant. By ECV-Maxist lights, the relevant circumstances include, roughly speaking, all and only the facts that are counterfactually independent of your decision, and A's objective value is its counterfactual value. By ERV-Maxist lights, by contrast, the relevant circumstances include all and only the unaffected facts and A's objective value equals RV(A). Proponents of both views can say that you *objectively* ought to choose A iff A is the option with the highest objective value or (in cases where two or more options are drawn for having the highest objective value) A is one of the options with the highest objective value. Moreover, both ECV-Maxists and ERV-Maxists can say that rational agents aim to maximize objective value by maximizing expected objective value. Thus, for ECV-Maxists, the choiceworthiness or utility (or, if you like, the *subjective value*) of an option equals

its expected counterfactual value (ECV), while for ERV-Maxists, it equals its expected real value (ERV). You *subjectively* ought to choose one of the options with the highest utility.

The argument from Fixity and U*-Max given earlier in this section was intended to provide some initial motivation for exploring the thesis that objective value equals real value rather than counterfactual value. However, the most important question to ask in deciding between the views is which of the two quantities satisfies condition (17). In other words, we need to ask whether facts about what it is rational to do in a given decision situation can be better explained by assuming that rational agents aim to maximize counterfactual value or by assuming that they aim to maximize real value. I will argue that the latter assumption allows for a better explanation of the data.

6. Generalizing the account

Suppose that ♦A is false. Then no possible A-world matches actuality perfectly in the unaffected facts, so that RV(A) does not exist. Consequently, if you are not certain of ♦A, then you are not certain that RV(A) exists, so that ERV(A) is undefined and ERV-Max falls silent about A's utility. To deal with such cases, we need to generalize ERV-Max. But before tackling this task (§6.2), we need to get clearer about the truth-conditions of counterfactuals.

6.1 Counterfactuals and causation

The account sketched in §2 entails that, for any option A, the criteria for an A-world *w*'s closeness to actuality include:

(a) *w*'s conformity to the actual laws.
(b) Match in pre-*t* facts (except in certain backwards-causation cases).

But what about post-*t* similarities? Do they also matter to the closeness of an A-world?

Some of them do while others do not. Consider a variant of an example due to Dorothy Edgington (2003, 2011).[11] You are about to watch an indeterministic lottery draw on television when someone offers to sell you ticket number 17. You decline. As luck would have it, 17 wins. It seems true to say that you would have won if you had bought the ticket. But that presupposes the following.

If you had bought ticket 17, then 17 would still have won.

Now suppose that the lottery company has two qualitatively indistinguishable lottery machines, X and Y, that give the same chance to every possible outcome. They used machine X. Consider:

If the company had used Y, 17 would still have won.

[11] Similar examples (called "Morgenbesser cases") are discussed in Adams 1975: 132−3, Tichý 1976, Slote 1978, Bennett 2003, Schaffer 2004, Kment 2006, 2014: Chs. 8−9, Wasserman 2006.

That sounds false—if Y had been used, then 17 might or might not have won. We hold the outcome of the draw fixed when evaluating the first counterfactual but not when assessing the second. It seems very plausible that this difference is explained by our causal beliefs. Your decision about whether to buy the ticket is causally unconnected to the outcome (or so we think). That is why the outcome can be held fixed in the first example. By contrast, the decision to use machine X is part of the outcome's causal history. That is why the outcome cannot be held fixed in the second case.

This suggests that (a)–(b) should be supplemented with another criterion of closeness.

(c) Match in post-*t* facts that are causally unaffected by your decision.[12]

What (a)–(c) have in common is that they concern inter-world similarities in unaffected facts: the laws, past facts (when there is no backwards causation), and unaffected future facts. This seems to be the unifying theme in all criteria of closeness. (See Kment 2006, 2014: Chs. 8–9 for more detailed arguments for this conclusion.) Therefore:

Unaffected-Fact Maximization. For any option A, the closest A-worlds are those that match actuality most closely in the unaffected facts.

6.2 Conditional expected real value

ERV-Max should be replaced with a rule that meets condition (18).

(18) (a) The new rule agrees with ERV-Max when you are certain of ♦A.
(b) Unlike ERV-Max, the new rule determines U(A) even when you are not certain of ♦A.

Now, ERV-Max can be understood as being based on two assumptions:

(19) (a) A's objective value is RV(A).
(b) A's utility equals your expectation of A's objective value.

(19)(a)−(b) entails (20).

(20) U(A) = E(RV(A), Cr)

There are two obvious methods of minimally revising this equation to obtain a rule that satisfies (18). Method 1 rejects (19)(a) but keeps (19)(b). It retains Cr in (20) but replaces RV with the quantity Q most similar to RV that meets the following conditions: $E(Q(A), Cr) = E(RV(A), Cr)$ whenever you are certain of ♦A, but $E(Q(A), Cr)$ is defined even when you are not certain of ♦A. Method 2 rejects (19)(b) but essentially retains (19)(a).[13] It keeps RV in (20) but replaces Cr with the probability function p most similar to Cr that meets the following conditions: $E(RV(A), p) = E(RV(A), Cr)$ whenever you are certain of ♦A, but $E(RV(A), p)$ is defined even when you are not

[12] For detailed arguments for conclusions along these lines, see Bennett 2003, Edgington 2003, 2011, Schaffer 2004, Kment 2006, 2014: Chs. 8−9, Wasserman 2006.

[13] More precisely, it retains a qualified version of (19)(a): if A's objective value exists, it equals RV(A).

certain of ♦A. As we will see, ECV-Max can be understood as resulting from Method 1 of revising ERV-Max. I will argue for Method 2.

> *Method 1*: *ECV-Max*. A's objective value is not in general RV(A). Rather, it is the closest approximation to RV(A) that is still defined even when RV(A) is not. To define this quantity, we do *not* appeal to the possible A-worlds that *perfectly* match actuality in the unaffected facts, as we do when defining RV(A). Instead, we appeal to the possible A-worlds that *most closely approximate* such perfect match in the unaffected facts. According to Unaffected-Fact Maximization, these are all and only the closest possible A-worlds. The value of the outcome that obtains at these worlds is A's counterfactual value, CV(A). Thus, CV(A) is A's objective value. A's utility (subjective value) equals its expected objective value. Hence, U(A) = ECV(A).

ECV-Max satisfies (18). To see this, suppose ♦A holds. Then some possible A-worlds match actuality perfectly in the unaffected facts. By Unaffected-Fact Maximization, these are all and only the closest A-worlds. Hence, if ♦A holds, then the outcome at the closest A-worlds is the outcome at the A-worlds that match actuality perfectly in the unaffected facts, so that CV(A) = RV(A). Therefore:

(21) $\mathrm{ECV}(A \mid \Diamond A) = \mathrm{ERV}(A \mid \Diamond A)$

If you are certain of ♦A, then ECV(A) = ERV(A) and ECV-Max agrees with ERV-Max. However, unlike RV(A), CV(A) is also defined when ♦A is false. Consequently, ECV(A) is defined even when you are not certain of ♦A.

> *Method 2: CERV-Max.* Whenever an option A has an objective value, that objective value is RV(A). Hence, A has an objective value only if ♦A holds. However, A's utility does not generally equal your *unconditional* expectation of A's objective value, but instead equals your expectation of A's objective value on the assumption that A's objective value exists. Thus, $\mathrm{U}(A) = \mathrm{E}(\mathrm{RV}(A), \mathrm{Cr}(- \mid \Diamond A)) = \mathrm{ERV}(A \mid \Diamond A)$. I will call $\mathrm{ERV}(A \mid \Diamond A)$ the *Conditional Expected Real Value* of A, or *CERV(A)*. ERV-Max should be replaced with CERV-Max (pronounced "serve-max").

> *CERV-Max*. U(A) = CERV(A)

Like ECV-Max, CERV-Max satisfies (18). If you are certain of ♦A, then CERV(A) = ERV(A | ♦A) = ERV(A), so that CERV-Max agrees with ERV-Max. However, CERV(A) is defined even when you are not certain of ♦A.[14,15]

It will be instructive to compare the procedures by which a probability function that is fit to play the role of Cr^A can be obtained from Cr according to ECV-Max and CERV-Max.

ECV-Max. ECV-Max entails (2): a probability function p is fit for the role of Cr^A iff $p(O) = Cr(A \boxright O)$ for all outcomes O. ECV-Maxists can obtain such a probability function from Cr by a variant (which I will call "ECV-imaging") of a procedure called *imaging Cr on A* (Lewis 1976, Joyce: Chs. 5–6). (See Figure 1a.) ECV-imaging shifts the probability that Cr assigns to any ¬A-world $w \in \Omega_{Cr}$ to some of the A-worlds closest to w, but does not move the probability Cr assigns to A-worlds.[16] Now, if $w \in \Omega_{Cr} \cap \langle \blacklozenge A \rangle$, then the background state b that contains w also contains A-worlds. The A-worlds in b are exactly those that match w perfectly in the unaffected facts. By Unaffected-Fact Match, they are exactly the A-worlds closest to w. ECV-imaging on A therefore moves w's probability to A-worlds in the same background state as w (as represented by the short arrows in Figure 1a). By contrast, if $w \in \Omega_{Cr} \cap \langle \neg \blacklozenge A \rangle$, then w is in some background state b that contains no A-worlds. ECV-imaging on A then moves w's probability outside of b, to the A-worlds that come closest to making b true (as represented by the long arrow). As Figure 1a shows, these A-worlds might be in epistemically impossible background states.

CERV-Max. By (21), CERV(A) = ERV(A | ♦A) = ECV(A | ♦A). Hence, CERV-Max entails that $U(A) = ECV(A \mid \blacklozenge A) = \Sigma_O \, Cr(A \boxright O \mid \blacklozenge A) \, V(O)$. The probability function Cr^A in (1) must therefore be chosen so that $Cr^A(O) = Cr(A \boxright O \mid \blacklozenge A)$. We can obtain a probability function satisfying

---

[14] Incompatibilits might believe that the truth of ♦A is required for you to be free to do A, or for you to have full control over what you do. However, it is important to note that this idea plays no role in the motivation for Method 2 that I outlined. The reason why you should evaluate A conditional on ♦A is not that ♦A's truth is a condition for being free to do A (or for having full control over what you do) and that you should assess A on the assumption that you are free to do A (or on the assumption that you have full control over what you do). The reason is rather that ♦A's truth is a condition for A's objective value to exist, and that you should assess A in light of your estimate of A's objective value conditional on the existence of A's objective value.

[15] ERV-Max resembles Brian Skyrms's theory, according to which $U(A) = \Sigma_{C,K} \, Cr(K) \, Cr(C \mid A \& K) \, V(C \& A \& K)$, where K ranges over "maximally specific specifications of the factors outside [the agent's] influence … which are causally relevant to the outcome" and C ranges over "specifications of factors which may be influenced by" the agent's action (Skyrms 1980: 133). Aside from motivation, the most important difference between the accounts is that the background states figuring in ERV-Max are maximally specific ways that *all* unaffected facts (not just of those causally relevant to the outcome) could be. *CERV-Max* differs much more significantly from Skyrms's account by making A's assessment conditional on ♦A.

[16] Suppose that probability function p results from ECV-imaging Cr on A. Then, $p(O) = Cr(A \boxright O)$. *Proof*. Note first that (31) holds for all $w \in \Omega_{Cr}$.

(31) ECV-imaging Cr on A shifts the probability Cr assigns to w to some of the A-worlds closest to w.

Proof of (31). If $w \in \Omega_{Cr} \cap \langle \neg A \rangle$, then (31) follows from my definition of ECV-imaging. If $w \in \Omega_{Cr} \cap \langle A \rangle$, then Unaffected-Fact Maximization entails that w is among the A-worlds closest to w (for, no A-world matches w more closely in unaffected facts than w). Now, if $w \in \Omega_{Cr} \cap \langle A \rangle$, then ECV-imaging Cr on A leaves the probability that Cr assigns to w on w, thereby in effect "shifting" w's probability to one of the A-worlds closest to w (viz., to w).

ECV-Max's underlying assumption (3) entails that, for any $w \in \Omega_{Cr}$, the same outcome obtains at all A-worlds closest to w. Hence, $A \boxright O$ holds at w (i.e., O holds at *all* the A-worlds closest to w) iff O holds at *any* of the A-worlds closest to w. Therefore, by (31), $A \boxright O$ holds at w iff O holds at the worlds to which ECV-imaging Cr on A shifts w's probability. Hence, $p(O)$ equals the sum of the probabilities that Cr assigns to $(A \boxright O)$-worlds. Therefore, $p(O) = Cr(A \boxright O)$.

this constraint from Cr in two steps (Figure 1b). *First*, we condition Cr on ♦A, which removes ¬♦A-worlds from the sample space. (⟨¬♦A⟩ is therefore blackened out in Figure 1b.) *Secondly*, we ECV-image Cr(− | ♦A) on A. Since $\Omega_{Cr(-|\Diamond A)} \subseteq \Diamond A$, ECV-imaging Cr(− | ♦A) on A shifts the probability of every ¬A-world $w \in \Omega_{Cr(-|\Diamond A)}$ to the A-worlds in the same background state as *w*. Consequently, the resulting probability function assigns no probability to epistemically impossible background states.

There are three closely interconnected reasons why ERV-Maxists should prefer Method 2 of revising their view to Method 1. *First*, Method 1 rests on the claim (22).

(22) A's objective value equals A's counterfactual value if ♦A is false.

It is unclear what could motivate (22). Remember the ERV-Maxist's rationale for identifying A's objective value with RV(A) in cases where ♦A holds: in such cases, RV(A) is the utility that You* assigns to A. To motivate (22) by a similar strategy, proponents of Method 1 would have to argue for (23).

(23) If ♦A is false, then CV(A) is the utility You* assigns to A.

However, it seems doubtful that You* assigns *any* utility to A if ¬♦A holds. If ¬♦A is true, then some unaffected facts PP are incompossible with A. You* is certain that PP are unaffected facts. To assign any utility to A, You* would need to consider possible A-worlds where some of PP fail to obtain, thereby violating Fixity. ((23) entails that You* commits such a violation of Fixity.) As mentioned in §4, it seems implausible that the right way to assess an option could ever involve a violation of Fixity.

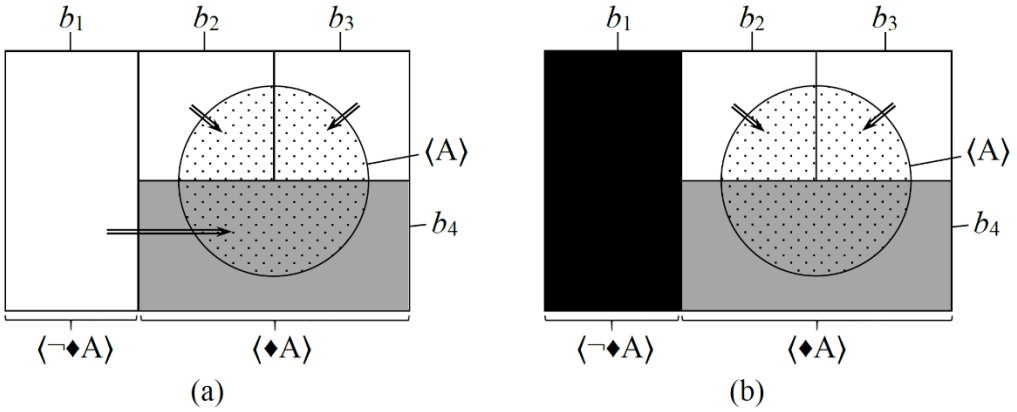

Figure 1. *Transforming Cr into Cr^A according to (a) ECV-Max and (b) CERV-Max.* Each diagram represents the set of all possible worlds. b_1–b_4 are the background states, b_1–b_3 are epistemically possible, b_2–b_4 are metaphysically compossible with A. The three unshaded (white) cells of diagram (a)—b_1, b_2, and b_3—represents Ω_{Cr}, those of diagram (b)—b_2 and b_3—represents $\Omega_{Cr(-|\Diamond A)}$.

Secondly, as we saw in §4, ECV-Max's predictions can contradict Fixity even for agents who do not believe ¬♦A. That is so because the procedure by which Cr can be transformed into a

probability function fit to play the role of Cr^A according to ECV-Max can shift probability to an epistemically impossible background state *b*. There might then be some proposition P that you take with certainty to be an unaffected fact but which fails to hold at *b*-worlds. In such cases, ECV-Maxists have to contradict Fixity by saying that $\Omega_{Cr^A} \not\subseteq \langle P \rangle$. By contrast, the process by which Cr can be transformed into a probability function fit to play the role of Cr^A according to CERV-Max only ever moves probability to epistemically possible background states. If you are certain that P is an unaffected fact, then P is true at all epistemically possible background states. Hence, by CERV-Max's lights, $\Omega_{Cr^A} \subseteq \langle P \rangle$. CERV-Max therefore never violates Fixity. Given the implausibility of such violations, this is an advantage of CERV-Max.

Thirdly, due to its violations of Fixity, ECV-Max gets examples like Past-Bet_1 and Law-Bet_1 wrong. To see that CERV-Max gets them right, consider the two examples in turn.

Past-Bet_1. We saw in §4 that Fixity entails the correct verdict about this example. Since CERV-Max entails Fixity, CERV-Max gets the case right.

Law-Bet_1. Ahmed claims that $U(Assert_L) > U(Assert_{\neg L})$ and shows that ECV-Max (combined with the non-backtracking account) entails that $U(Assert_{\neg L}) \geq U(Assert_L)$. Now, CERV-Max does *not* entail that $U(Assert_{\neg L}) \geq U(Assert_L)$. By CERV-Max, $U(Assert_L) = Cr(L \mid \blacklozenge Assert_L)$ and $U(Assert_{\neg L}) = Cr(\neg L \mid \blacklozenge Assert_{\neg L})$. [17] Thus, CERV-Max agrees with Ahmed's claim that $Assert_L$ is uniquely rational, provided that (24) holds.

(24) $Cr(L \mid \blacklozenge Assert_L) > Cr(\neg L \mid \blacklozenge Assert_{\neg L})$

Since $Cr(L) > Cr(\neg L)$, (24) is true if L is probabilistically independent of $\blacklozenge Assert_L$ and $\blacklozenge Assert_{\neg L}$.

However, CERV-Max *disagrees* with Ahmed if (24) is false. Consider:

> *Law-Bet_3*. This example is like Law-Bet_1, except that you are in a room with three blackboards, B_1, B_2, and B_3, on which are written Q_1, Q_2, and Q_3, respectively (see below). You are certain that the message on exactly one blackboard is true, that a random device was used to decide which blackboard would carry the true message, and that B_1, B_2, and B_3 had chances 60%, 10%, and 30%, respectively, of being picked. Therefore, $Cr(Q_1) = .6$, $Cr(Q_2) = .1$, $Cr(Q_3) = .3$. Moreover, assume that $Cr(Assert_L \mid Q_3) = \frac{1}{3}$. You are certain that there is no backwards causation.
>
> Q_1: L holds. L and the state of the universe yesterday metaphysically necessitate $Assert_{\neg L}$.
> Q_2: L holds. L and the state of the universe yesterday metaphysically necessitate $Assert_L$.
> Q_3: $\neg L$ holds. The laws are indeterministic. $Assert_L$'s present physical chance equals $\frac{1}{3}$.

[17] As we saw at the end of §4, it follows from Fixity (and therefore from CERV-Max) that $U(Assert_L \mid L) = 1$, and $U(Assert_L \mid \neg L) = 0$. By CERV-Max and the law of total expectation:

$$\begin{aligned} U(Assert_L) &= ERV(Assert_L \mid \blacklozenge Assert_L) \\ &= Cr(L \mid \blacklozenge Assert_L)\ ERV(Assert_L \mid \blacklozenge Assert_L \& L) + Cr(\neg L \mid \blacklozenge Assert_L)\ ERV(Assert_L \mid \blacklozenge Assert_L \& \neg L) \\ &= Cr(L \mid \blacklozenge Assert_L)\ U(Assert_L \mid L) + Cr(\neg L \mid \blacklozenge Assert_L)\ U(Assert_L \mid \neg L) \\ &= Cr(L \mid \blacklozenge Assert_L) \end{aligned}$$

By analogous reasoning, $U(Assert_{\neg L}) = Cr(\neg L \mid \blacklozenge Assert_{\neg L})$.

The options are probabilistically independent of L.[18] However, $\mathrm{CERV}(\mathrm{Assert}_{\neg L}) = \mathrm{Cr}(\neg L \mid \blacklozenge \mathrm{Assert}_{\neg L}) = \frac{1}{3} > \frac{1}{4} = \mathrm{Cr}(L \mid \blacklozenge \mathrm{Assert}_L) = \mathrm{CERV}(\mathrm{Assert}_L)$.[19] By CERV-Max, you should assert ¬L.

I think that this verdict is plausible on reflection. To evaluate Assert_L, you need to determine the probability of True under the supposition Assert_L. Since Q_1 entails $\mathrm{Assert}_{\neg L}$, $\mathrm{Cr}^{\mathrm{Assert}_L}(Q_1) = 0$. By Past−Law Unaffectedness, you are certain that $Q_1 \vee Q_2 \vee Q_3$ is an unaffected fact. It follows by Fixity that $\mathrm{Cr}^{\mathrm{Assert}_L}(Q_1 \vee Q_2 \vee Q_3) = 1$ and therefore that $\mathrm{Cr}^{\mathrm{Assert}_L}(Q_2 \vee Q_3) = 1$. We have reason to accept this conclusion, since Fixity is plausible. Now, Blackboard B_3 had three times as high a chance of being picked to carry a true message as B_2, so $\mathrm{Cr}(Q_3) / \mathrm{Cr}(Q_2) = 3$. For those who reject EDT's predictions about Newcomb, there is no obvious reason for thinking that $\mathrm{Cr}^{\mathrm{Assert}_L}(Q_3) / \mathrm{Cr}^{\mathrm{Assert}_L}(Q_2)$ differs from $\mathrm{Cr}(Q_3) / \mathrm{Cr}(Q_2)$. After all, by Past−Law Unaffecteness, Q_2 and Q_3 are causally unaffected by your decision. Moreover, conditioning on $\blacklozenge\mathrm{Assert}_L$ leaves the ratio of their probabilities unchanged.[20] (It is true that the ratio of their probabilities conditional on Assert_L differs from the ratio of their unconditional probabilities.[21] But this reflects purely epistemic, non-causal dependencies of Q_2 and Q_3 on Assert_L. For those who disagree with EDT about Newcomb, such dependencies should seem irrelevant to the probabilities of Q_2 and Q_3 under the supposition Assert_L.) Assuming that $\mathrm{Cr}^{\mathrm{Assert}_L}(Q_3) / \mathrm{Cr}^{\mathrm{Assert}_L}(Q_2) = \mathrm{Cr}(Q_3) / \mathrm{Cr}(Q_2) = 3$, it follows that $\mathrm{U}(\mathrm{Assert}_L) = \mathrm{Cr}^{\mathrm{Assert}_L}(L) = \mathrm{Cr}^{\mathrm{Assert}_L}(Q_2) = \mathrm{Cr}^{\mathrm{Assert}_L}(Q_2 \mid Q_2 \vee Q_3) = \mathrm{Cr}^{\mathrm{Assert}_L}(Q_2) / (\mathrm{Cr}^{\mathrm{Assert}_L}(Q_2) + \mathrm{Cr}^{\mathrm{Assert}_L}(Q_3)) = \mathrm{Cr}^{\mathrm{Assert}_L}(Q_2) / (\mathrm{Cr}^{\mathrm{Assert}_L}(Q_2) + 3\,\mathrm{Cr}^{\mathrm{Assert}_L}(Q_2)) = \frac{1}{4}$. By analogous reasoning, $\mathrm{U}(\mathrm{Assert}_{\neg L}) = \frac{1}{3}$.[22] Therefore, $\mathrm{U}(\mathrm{Assert}_{\neg L}) > \mathrm{U}(\mathrm{Assert}_L)$.

One might argue against CERV-Max's prediction on the grounds that it violates the following prima facie plausible principle (Ahmed 2013b: 292, formulation simplified).

> *Causal Betting Principle (CBP)*: If you must choose between Bet_Q on Q and Bet_R on R, if the bets have the same payoffs for winning and losing, if you are certain that Q and R are unaffected, and if $\mathrm{Cr}(Q) > \mathrm{Cr}(R)$, then Bet_Q is uniquely rational.

This argument is unsound, however, since CBP is refuted by Past-Bet_2.

> *Past-Bet₂*. Arm counts as betting on P_1, ¬Arm as betting on $\neg P_1$. The two bets have the same payoffs for winning and for losing. $\mathrm{Cr}(P_1) > .5$. You are certain that P_1 is unaffected.
>
> P_1: The state of the universe yesterday at noon, combined with the natural laws of the actual world, metaphysically necessitates ¬Arm.

CBP predicts that you should choose Arm. But that seems wrong, since choosing Arm is self-undermining. CERV-Max tells us how to revise CBP, since it entails that "Cr(Q) > Cr(R)" in CBP

---

[18] $\mathrm{Cr}(\mathrm{Assert}_L \mid L) = \mathrm{Cr}(Q_2) / (\mathrm{Cr}(Q_1) + \mathrm{Cr}(Q_2)) = \frac{1}{2}$, $\mathrm{Cr}(\mathrm{Assert}_L \mid \neg L) = \mathrm{Cr}(\mathrm{Assert}_L \mid Q_3) = \frac{1}{2}$.

[19] $\mathrm{Cr}(L \mid \blacklozenge\mathrm{Assert}_L) = \mathrm{Cr}(L \,\&\, \blacklozenge\mathrm{Assert}_L) / \mathrm{Cr}(\blacklozenge\mathrm{Assert}_L) = \mathrm{Cr}(Q_2) / (\mathrm{Cr}(Q_2) + \mathrm{Cr}(Q_3)) = \frac{1}{4}$
$\mathrm{Cr}(\neg L \mid \blacklozenge\mathrm{Assert}_{\neg L}) = \mathrm{Cr}(\neg L \,\&\, \blacklozenge\mathrm{Assert}_{\neg L}) / \mathrm{Cr}(\blacklozenge\mathrm{Assert}_{\neg L}) = \mathrm{Cr}(Q_3) / (\mathrm{Cr}(Q_1) + \mathrm{Cr}(Q_3)) = \frac{1}{3}$

[20] $\mathrm{Cr}(Q_3 \mid \blacklozenge\mathrm{Assert}_L) = \mathrm{Cr}(Q_3) / (\mathrm{Cr}(Q_2) + \mathrm{Cr}(Q_3)) = \frac{3}{4}$, $\mathrm{Cr}(Q_2 \mid \blacklozenge\mathrm{Assert}_L) = \mathrm{Cr}(Q_2) / (\mathrm{Cr}(Q_2) + \mathrm{Cr}(Q_3)) = \frac{1}{4}$, so $\mathrm{Cr}(Q_3 \mid \blacklozenge\mathrm{Assert}_L) / \mathrm{Cr}(Q_2 \mid \blacklozenge\mathrm{Assert}_L) = 3 = \mathrm{Cr}(Q_3) / \mathrm{Cr}(Q_2)$.

[21] $\mathrm{Cr}(Q_3 \mid \mathrm{Assert}_L) = \mathrm{Cr}(Q_3 \,\&\, \mathrm{Assert}_L) / \mathrm{Cr}(\mathrm{Assert}_L) = \mathrm{Cr}(Q_3 \,\&\, \mathrm{Assert}_L) / (\mathrm{Cr}(Q_2) + \mathrm{Cr}(Q_3 \,\&\, \mathrm{Assert}_L)) = (.3/7) / (.1 + (.3/7)) = .3$. $\mathrm{Cr}(Q_2 \mid \mathrm{Assert}_L) = \mathrm{Cr}(Q_2 \,\&\, \mathrm{Assert}_L) / (\mathrm{Cr}(Q_2) + \mathrm{Cr}(Q_3 \,\&\, \mathrm{Assert}_L)) = .1 / (.1 + (.3/7)) = .7$. $\mathrm{Cr}(Q_3 \mid \mathrm{Assert}_L) / \mathrm{Cr}(Q_2 \mid \mathrm{Assert}_L) = \frac{3}{7} \neq 3 = \mathrm{Cr}(Q_3) / \mathrm{Cr}(Q_2)$.

[22] $\mathrm{Cr}^{\mathrm{Assert}_{\neg L}}(Q_1 \vee Q_3) = 1$ and $\mathrm{Cr}^{\mathrm{Assert}_{\neg L}}(Q_1) / \mathrm{Cr}^{\mathrm{Assert}_{\neg L}}(Q_3) = \mathrm{Cr}(Q_1) / \mathrm{Cr}(Q_3) = 2$, so $\mathrm{U}(\mathrm{Assert}_{\neg L}) = \mathrm{Cr}^{\mathrm{Assert}_{\neg L}}(\neg L) = \frac{1}{3}$.

needs to be replaced with "$Cr(Q \mid \blacklozenge Bet_Q) > Cr(R \mid \blacklozenge Bet_R)$". In Past-Bet$_2$, $Cr(P_1 \mid \blacklozenge Arm) = 0 \leq Cr(\neg P_1 \mid \blacklozenge \neg Arm)$. The revised version of CBP therefore does not entail that Arm is uniquely rational. It also entails that you should choose $Assert_{\neg L}$ in Law-Bet$_3$, since $Cr(\neg L \mid \blacklozenge Assert_{\neg L}) = \frac{1}{3} > \frac{1}{4} = Cr(L \mid \blacklozenge Assert_L)$.

7. Collapse, and CERV-Max as CDT–EDT-hybrid

CERV-Max collapses into CDT in some special cases, into EDT in others. More precisely:

Det_A: The unaffected facts either (metaphysically) necessitate A or (metaphysically) necessitate ¬A.

CDT-Collapse. If you are certain of ♦A, then CERV(A) = ECV(A).[23]

EDT-Collapse. If you are certain of Det_A, then CERV(A) = EV(A).[24]

Past-Bet$_1$ illustrates EDT-Collapse: you are certain of Det_{Arm}, and CERV-Max and EDT assign the same utilities. §8.1 will describe an example of CDT-Collapse. Since $\neg Det_A$ entails ♦A, (25) follows from CDT-Collapse.

(25) If you are certain of $\neg Det_A$, then CERV(A) = ECV(A).

CDT-Collapse and EDT-Collapse deal with special cases. But we can use them to derive another result that applies more widely.

Hybrid. Provided all relevant terms are defined,

$U(A) = Cr(Det_A \mid \blacklozenge A)\, EV(A \mid Det_A) + Cr(\neg Det_A \mid \blacklozenge A)\, ECV(A \mid \neg Det_A)$.

Proof. By EDT-Collapse, $CERV(A \mid Det_A) = EV(A \mid Det_A)$. By (25), $CERV(A \mid \neg Det_A) = ECV(A \mid \neg Det_A)$. Therefore ("LTE" abbreviates "Law of total expectation"):

---

[23] *Proof*. By (21), $CERV(A) =_{def} ERV(A \mid \blacklozenge A) = ECV(A \mid \blacklozenge A)$. If you are certain of ♦A, then $ECV(A \mid \blacklozenge A) = ECV(A)$.

[24] *Proof*. Suppose you are certain of Det_A. Det_A metaphysically necessitates $A \leftrightarrow \blacklozenge A$. Hence:

(32) $Cr(- \mid \blacklozenge A) = Cr(- \mid A)$

Let $\mathcal{B}_{\blacklozenge A} =_{def} \{b \in \mathcal{B}: b \cap \langle A \rangle \neq \emptyset\} = \{b \in \mathcal{B}: b \cap \langle \blacklozenge A \rangle \neq \emptyset\} = \{b \in \mathcal{B}: b \subseteq \langle \blacklozenge A \rangle\}$. Note that:

(33) Conditional on ♦A, you are certain of $\cup \mathcal{B}_{\blacklozenge A}$.

By Predetermined Outcome, for every $b \in \mathcal{B}_{\blacklozenge A}$, $A \,\&\, b$ necessitates a unique outcome $O_{A\&b}$. Hence:

(34) For all $b \in \mathcal{B}_{\blacklozenge A}$, $Cr(O_{A\&b} \mid A \,\&\, b) = 1$, and $Cr(O \mid A \,\&\, b) = 0$ for all $O \neq O_{A\&b}$.

$EV(A) = \sum_O Cr(O|A)\, V(O)$

$= \sum_O Cr(O|\blacklozenge A)\, V(O)$ (32)

$= \sum_O \left(\sum_{b \in \mathcal{B}_{\blacklozenge A}} Cr(b|\blacklozenge A)\, Cr(O|\blacklozenge A \,\&\, b)\right) V(O)$ (33), Law of total probability

$= \sum_O \left(\sum_{b \in \mathcal{B}_{\blacklozenge A}} Cr(b|\blacklozenge A)\, Cr(O|A \,\&\, b)\right) V(O)$ (32)

$= \sum_{b \in \mathcal{B}_{\blacklozenge A}} Cr(b \mid \blacklozenge A)\, V(O_{A\&b})$ (34)

$= ERV(A|\blacklozenge A)$ (15), (33)

$= CERV(A)$

$$
\begin{aligned}
U(A) &= ERV(A \mid \blacklozenge A) && \text{CERV-Max}\\
&= Cr(Det_A \mid \blacklozenge A)\, ERV(A \mid \blacklozenge A \,\&\, Det_A) &&\\
&\quad + Cr(\neg Det_A \mid \blacklozenge A)\, ERV(A \mid \blacklozenge A \,\&\, \neg Det_A) && \text{LTE}\\
&= Cr(Det_A \mid \blacklozenge A)\, CERV(A \mid Det_A) + Cr(\neg Det_A \mid \blacklozenge A)\, CERV(A \mid \neg Det_A) &&\\
&= Cr(Det_A \mid \blacklozenge A)\, EV(A \mid Det_A) + Cr(\neg Det_A \mid \blacklozenge A)\, ECV(A \mid \neg Det_A) &&
\end{aligned}
$$

According to Hybrid, you can evaluate A as follows. First, you compute A's utility on the assumption that the facts beyond your influence determine whether you do A (Det_A). You do so by applying EDT. Next, you calculate A's utility on the opposite assumption. This can be done by applying ECV-Max. Finally, you compute U(A) as the $Cr(- \mid \blacklozenge A)$-weighted average of the two conditional utilities. A's utility is thus a composite of an evidential component and a counterfactual or causal component.

8. Newcomb and dominance reconsidered

8.1 CERV-Max and Newcomb

Let a *Newcomb scenario* be any example that is a version of the case labeled "Newcomb" in §1, except that your confidence in the oracle's prediction need not be exactly 99% but might have some other very high value. (Thus, there are Newcomb scenarios in which your confidence is 90%, and others in which you are certain that the prediction is correct.) CDT is motivated by the thought that two-boxing is uniquely rational in any Newcomb scenario. However, if Fixity holds, then that is not true. Consider:

> *Pre-Determined Newcomb*. Like Newcomb of §1, except that you are certain of the following: some unaffected facts (including certain laws and facts about the past) necessitate that the oracle's prediction is correct and that $B_1 \supset s_M$ and $B_2 \supset s_0$ hold.

By Guaranteed Outcome (and hence by Fixity), $U(B_1) = 1{,}000{,}000 > U(B_2) = 1{,}000$.

This verdict is perhaps not surprising. In *Pre-Determined Newcomb*, you are *certain* that the oracle is correct. As some philosophers have noted, you might be tempted to think that you should one-box in such cases, even if you believe that two-boxing is rational in other Newcomb scenarios (Nozick 1969: 140–1, Levi 1975; also see Seidenfeld 1984, Sobel 1988).

According to CERV-Max, however, what it is rational to do in a Newcomb scenario is not generally determined by whether you are certain that the oracle is right. Instead, it depends on your attitude towards Det.

> Det: The unaffected facts either necessitate B_1 or necessitate B_2.

Consider three types of Newcomb scenario.

> *Newcomb*$_1$: You are certain of Det.
>
> *Newcomb*$_2$: You are certain neither of Det nor of ¬Det.
>
> *Newcomb*$_3$: You are certain of ¬Det.

In Newcomb$_1$ you might or might not be certain that the prediction is correct. Irrespective of that, CERV-Max and EDT-Collapse entail that one-boxing is uniquely rational. In Newcomb$_3$, you are certain of both $\blacklozenge B_1$ and $\blacklozenge B_2$, so that, by CERV-Max and CDT-Collapse, you should two-box. Finally, in Newcomb$_2$ the utilities of your options are determined in accordance with Hybrid. It depends on the details of your credence function whether one-boxing or two-boxing is rational.

8.2 CERV-Max and dominance

The most influential argument for two-boxing is the Dominance Argument of §1. Its underlying dominance principle can be stated as follows:

> For any options A and B, H is an *(A, B)-partition* iff$_{\text{def}}$ H is a countable set of pairwise metaphysically incompossible propositions such that:
> (a) $\{\langle h\rangle \cap \Omega_{Cr}: h \in H\}$ partitions Ω_{Cr} (i.e., every $h \in H$ is epistemically possible and you are certain that some $h \in H$ is true).
> (b) For every $h \in H$, $A \& h$ necessitates a unique outcome O_{Ah}, and $B \& h$ necessitates a unique outcome O_{Bh}.

(26) *Dominance*. $U(A) > U(B)$ if there is an (A, B)-partition H such that:
  (a) You are certain that the truth-value of each $h \in H$ is counterfactually independent of A and of B.
  (b) For all $h \in H$, $V(O_{Ah}) \geq V(O_{Bh})$.
  (c) For some $h \in H$, $Cr(h) > 0$ and $V(O_{Ah}) > V(O_{Bh})$.

In any Newcomb scenario, $\{s_0, s_M\}$ is a (B_1, B_2)-partition. Since $V(O_{B_2 s_0}) > V(O_{B_1 s_0})$ and $V(O_{B_2 s_M}) > V(O_{B_1 s_M})$, Dominance entails that $U(B_2) > U(B_1)$.

This consequence of Dominance—that two-boxing is rational in *all* Newcomb cases—contradicts CERV-Max. CERV-Maxists therefore need to reject Dominance. Past-Bet$_1$ provides independent reasons for this move, since Dominance yields the wrong outcome in Past-Bet$_1$. In Past-Bet$_1$, you are certain that (27) is both true and counterfactually independent of your choice.

(27) $((\text{Arm} \,\&\, P_1) \supset \$10) \,\&\, ((\text{Arm} \,\&\, \neg P_1) \supset -\$1) \,\&$
$((\neg\text{Arm} \,\&\, P_1) \supset \$1) \,\&\, ((\neg\text{Arm} \,\&\, \neg P_1) \supset -\$10)$

You are also certain that P_1 is counterfactually independent of your choice. Therefore, $\{P_1 \,\&\, (27), \neg P_1 \,\&\, (27)\}$ is an (Arm, ¬Arm)-partition that meets condition (26)(a). It also satisfies (26)(b) and (c), since $V(O_{\text{Arm}, P_1 \,\&\, (27)}) = 10 > 1 = V(O_{\neg\text{Arm}, P_1 \,\&\, (27)})$ and $V(O_{\text{Arm}, \neg P_1 \,\&\, (27)}) = -1 > -10 = V(O_{\neg\text{Arm}, \neg P_1 \,\&\, (27)})$. Dominance therefore yields the wrong result that $U(\text{Arm}) > U(\neg\text{Arm})$.

Dominance could of course be revised in different ways. However, the most natural revision starts from the observation that Dominance is an instance of the following schematic principle (where "X" is a placeholder).

Schematic Dominance. $U(A) > U(B)$ if there is an (A, B)-partition H that meets conditions X, (26)(b), and (26)(c).

If you endorse a decision theory of the form (1), you can transform Schematic Dominance into a principle that is entailed by your decision theory as follows. First, pick some condition C on (A, B)-partitions for which your theory entails that (28) is true.

(28) If an (A, B)-partition H satisfies C, then $Cr^A(h) = Cr^B(h) = Cr(h)$ for all $h \in H$.

Next, replace "X" in Schematic Dominance with a name for condition C. It can be shown that the resulting principle is true according to your theory.[25]

Which conditions C you take to satisfy (28) depends on which probability functions you take to be fit to play the roles of Cr^A and Cr^B. For ECV-Maxists, who believe that the probability distribution obtained from Cr by ECV-imaging on A (B) can play the role of Cr^A (Cr^B), (26)(a) states a condition C that satisfies (28).[26] That is why their view validates Dominance. For evidentialists, who believe that $Cr^A(O) = Cr(O|A)$ and $Cr^B(O) = Cr(O|B)$, the following condition satisfies (28): for all $h \in H$, $Cr(h|A) = Cr(h|B) = Cr(h)$. For CERV-Maxists, who believe that a probability function fit to play the role of Cr^A (Cr^B) can be obtained from $Cr(-|\blacklozenge A)$ ($Cr(-|\blacklozenge B)$) by ECV-imaging on A (B), the condition (29) satisfies (28).[27]

(29) (a) You are certain that the truth-value of each $h \in H$ is counterfactually independent of A and of B, and
(b) for each $h \in H$, $Cr(h|\blacklozenge A) = Cr(h|\blacklozenge B) = Cr(h)$.

---

[25] Suppose that some (A,B)-partition H satisfies the conditions C and (26)(b)–(c), and assume that (1) and (28) are true. Note that:

(35) For any $h \in H$ and outcome O: $Cr^A(O|h) = 1$ if $O = O_{Ah}$, $Cr^A(O|h) = 0$ otherwise.

Hence:

(36) $U(A) = \Sigma_O\, Cr^A(O)\, V(O)$ (1)
$= \Sigma_{h \in H, O}\, Cr^A(h)\, Cr^A(O|h)\, V(O)$
$= \Sigma_{h \in H}\, Cr^A(h)\, V(O_{Ah})$ (35)
$= \Sigma_{h \in H}\, Cr(h)\, V(O_{Ah})$ (28), H satisfies condition C

By analogous reasoning:

(37) $U(B) = \Sigma_{h \in H}\, Cr(h)\, V(O_{Bh})$

From (36), (37), and the fact that H satisfies (26)(b)–(c), we can infer that $U(A) > U(B)$.

[26] If an (A,B)-partition H satisfies (26)(a) and the probability function P^A (P^B) results from ECV-imaging Cr on A (B), then $P^A(h) = P^B(h) = Cr(h)$ for all $h \in H$. The proof is a variant of the proof given in the next footnote.

[27] To see this, let H be an (A,B)-partition that satisfies condition (29). From (29)(a), we can infer (38).

(38) For any $w \in \Omega_{Cr} \cap \langle\blacklozenge A\rangle$, $h \in H$: $A \boxright h$ holds at w if h holds at w, and $A \boxright \neg h$ holds at w if $\neg h$ holds at w.

By CERV-Max, Cr^A can be obtained from $Cr(-|\blacklozenge A)$ by ECV-imaging on A. For any proposition P and any $w \in \Omega_{Cr} \cap \langle\blacklozenge A\rangle$, if $A \boxright P$ is true at w, then ECV-imaging $Cr(-|\blacklozenge A)$ on A shifts w's probability to $\langle P\rangle$-worlds. Given (38), it follows that, for any $h \in H$, ECV-imaging $Cr(-|\blacklozenge A)$ on A shifts w's probability to $\langle h\rangle$-worlds if h is true at w and to $\langle \neg h\rangle$-worlds if $\neg h$ is true at w. Hence, $Cr^A(h) = Cr(h|\blacklozenge A)$. From this and (29)(b), we can infer that $Cr^A(h) = Cr(h)$. By analogous reasoning, $Cr^B(h) = Cr(h)$ for all $h \in H$. So, (29) is a condition C that satisfies (28).

Replacing "X" in Schematic Dominance with (29) yields the following principle:

Restricted Dominance. $U(A) > U(B)$ if there is an (A,B)-partition H such that:

(a) You are certain that the truth-value of each $h \in H$ is counterfactually independent of A and of B.

(b) For each $h \in H$, $Cr(h|\blacklozenge A) = Cr(h|\blacklozenge B) = Cr(h)$.

(c) For all $h \in H$, $V(O_{Ah}) \geq V(O_{Bh})$.

(d) For some $h \in H$, $Cr(h) > 0$ and $V(O_{Ah}) > V(O_{Bh})$.

Restricted Dominance cannot be applied to the (B_1, B_2)-partition $\{s_0, s_M\}$ in Newcomb$_1$, since $Cr(s_0|\blacklozenge B_1) \neq Cr(s_0|\blacklozenge B_2)$ and $Cr(s_M|\blacklozenge B_1) \neq Cr(s_M|\blacklozenge B_2)$.[28] For analogous reasons, the principle cannot generally be used in Newcomb$_2$. In Newcomb$_3$, however, $Cr(s_0|\blacklozenge B_1) = Cr(s_0|\blacklozenge B_2)$ and $Cr(s_M|\blacklozenge B_1) = Cr(s_M|\blacklozenge B_2)$,[29] so that we can infer from Restricted Dominance that two-boxing is uniquely rational.

9. The discontinuity objection

Suppose you are the agent in a Newcomb scenario, that Cr(The oracle's prediction is true) = .99, and that $Cr(s_0) = Cr(s_M) = .5$. Let t be a time immediately before your decision and let ch_t be the chance distribution at t. If you are certain of (30), then you are certain of ¬Det. CERV-Max then agrees with ECV-Max that $U(B_1) = 500{,}000$ and $U(B_2) = 501{,}000$. However, if you are certain of Det, then CERV-Max agrees with EDT that $U(B_1) = 990{,}000$ and $U(B_2) = 11{,}000$.

(30) The unaffected facts necessitate either $ch_t(B_1) = .9999$ or $ch_t(B_2) = .9999$, but they do not necessitate B_1 or B_2.

The difference between certainty in (30) and certainty in Det can seem like the difference between being certain that the unaffected facts *almost but not quite* determine your action and being certain that they *do* determine your action. Isn't it implausible that such a small difference in your credal state should make such a big difference to your utilities? To throw the problem into relief, let us imagine that you gradually increase your estimate of how close the unaffected facts come to determining your action. More precisely, you move from being certain that they necessitate either $ch_t(B_1) = .99$ or $ch_t(B_2) = .99$, to being certain that they necessitate either $ch_t(B_1) = .999$ or $ch_t(B_2) = .999$, etc., to finally being certain of Det. CERV-Max entails that $U(B_1)$ and $U(B_2)$ stay constant at 500,000 and 501,000 (respectively) throughout this process until you come to accept Det. Then the utilities suddenly change to 990,000 and 11,000, respectively. Isn't it implausible that there should be such an abrupt discontinuity (see Seidenfeld 1984, Ahmed 2015)?

---

[28] Let R (W) be the proposition that the oracle's prediction is right (wrong). In Newcomb$_1$, you are certain that Det holds and are therefore certain of $\blacklozenge B_1 \leftrightarrow B_1$ and $\blacklozenge B_2 \leftrightarrow B_2$, so that $Cr(-|\blacklozenge B_1) = Cr(-|B_1)$ and $Cr(-|\blacklozenge B_2) = Cr(-|B_2)$. Hence, $Cr(s_0|\blacklozenge B_1) = Cr(s_0|B_1) = Cr(W) \neq Cr(R) = Cr(s_0|B_2) = Cr(s_0|\blacklozenge B_2)$. Analogous reasoning shows that $Cr(s_M|\blacklozenge B_1) \neq Cr(s_M|\blacklozenge B_2)$.

[29] In Newcomb$_3$, you are certain of ¬Det and therefore of $\blacklozenge B_1$ and $\blacklozenge B_2$, so that $Cr(-|\blacklozenge B_1) = Cr(-|\blacklozenge B_2) = Cr$.

Reply. The discontinuity seems less surprising and implausible once we notice that its existence follows from two independently reasonable assumptions: (i) your utilities depend crucially on your credences concerning s_M and s_0; (ii) you should assess each option A conditional on ♦A. (i) is obvious, and we noted in §6.2 that (ii) is also plausible. Given (ii), what matters to your utilities are not your *unconditional* credences in s_M and s_0, but your credence in these propositions conditional on $\Diamond B_1$ and on $\Diamond B_2$. Now, these conditional credences change greatly and abruptly as you move from certainty in (30) to certainty in Det. For, if you are certain of (30), then you are certain of $\Diamond B_1$ and $\Diamond B_2$, so that $Cr(s_0 \mid \Diamond B_1) = Cr(s_0 \mid \Diamond B_2) = Cr(s_0) = .5$ and $Cr(s_M \mid \Diamond B_1) = Cr(s_M \mid \Diamond B_2) = Cr(s_M) = .5$. But if you are certain of Det, then you are certain of $\Diamond B_1 \leftrightarrow B_1$ and $\Diamond B_2 \leftrightarrow B_2$, so that $Cr(s_0 \mid \Diamond B_1) = Cr(s_0 \mid B_1) = .01$ and therefore $Cr(s_M \mid \Diamond B_1) = .99$, and $Cr(s_0 \mid \Diamond B_2) = Cr(s_0 \mid B_2) = .99$ and hence $Cr(s_M \mid \Diamond B_2) = .01$. Given this big and abrupt change in the propositional attitudes on which your utilities depend, it is unsurprising that your utilities also change greatly and abruptly.

10. Comparisons

I will conclude by discussing two alternative responses to Ahmed cases.[30]

Sandgren and Williamson (*S&W*). S&W (2020) propose a revision of ECV-Max to get Past-Bet$_1$ right. They argue that an option-to-outcome counterfactual $A \mathrel{\Box\!\!\rightarrow} O$ is "irrelevant for the purposes of deliberation when you are certain that making its consequent true given its antecedent would violate the actual laws" (4), since in such cases you can be sure that A is not a means to achieving O. Outcome O is then irrelevant to the evaluation of A, and $Cr^A(O)$ should equal 0 (rather than $Cr(A \mathrel{\Box\!\!\rightarrow} O)$, as ECV-Maxists assume). To renormalize Cr^A, $Cr^A(O^*)$ must be increased proportionally for all *relevant* outcomes O*. Thus, where $R_A =_{def} \cup\{\langle A \mathrel{\Box\!\!\rightarrow} O\rangle \mid O \text{ is relevant}\}$, $Cr^A(O) = Cr(A \mathrel{\Box\!\!\rightarrow} O \mid R_A)$. Therefore:

Sandgren-Williamson Theory$_1$ (*SWT$_1$*). $U(A) = \sum_O Cr(A \mathrel{\Box\!\!\rightarrow} O \mid R_A)\, V(O) = ECV(A \mid R_A)$

(S&W only partially specify the conditions for an outcome's relevance: certainty that A & O would involve a law-violation is sufficient for O's irrelevance to the evaluation of A, but S&W do not claim that it is necessary. Instead, they toy with, but do not endorse, the idea that knowing that A & O would involve a law-violation, or high confidence in this proposition, is also sufficient for O's irrelevance.) S&W argue that in examples like Past-Bet$_1$, the outcome of winning is irrelevant to the evaluation Arm and the outcome of losing is irrelevant to the assessment of ¬Arm. Hence, $R_{Arm} = \langle Arm \mathrel{\Box\!\!\rightarrow} -\$1\rangle$, $Cr^{Arm}(\$10) = 0$, $Cr^{Arm}(-\$1) = 1$, and $U(Arm) = -1$. By analogous reasoning, $U(\neg Arm) = 1$.

SWT$_1$'s main idea and motivation differ from CERV-Max's, since SWT$_1$ rests on Optimal Effect while CERV-Max relies on Fixity. Both views entail that $U(A) = ECV(A \mid s)$ for some state *s*. $s = R_A$ according to SWT$_1$, $s = \langle \Diamond A\rangle$ according to CERV-Max. When $R_A \neq \langle \Diamond A\rangle$, SWT$_1$ and CERV-

[30] Wedgwood's "benchmark theory" (Wedgwood 2013) also yields the right result in Ahmed cases. However, his view faces what appear to me to be counterexamples, including the case discussed in Wedgwood 2013: 2671ff. in response to an objection by Ray Briggs (I disagree with Wedgwood's verdict about this case) and examples presented in Bassett 2015: §4.1.

Max make different predictions. Law-Bet$_1$ is a case in point. CERV-Max gets this case right while SWT$_1$ does not (as S&W acknowledge in Williamson and Sandgren 2021). Williamson and Sandgren 2021 proposes two possible revisions of ECV-Max each of which they take to accommodate Law-Bet$_1$. *Proposal 1* uses an alternative account of counterfactuals (Nolan 2017): the A-worlds closest to a possible world *w* match *w* throughout the pre-antecedent period *and* perfectly conform to *w*'s laws, even if *w*'s laws and pre-antecedent history metaphysically necessitate ¬A (in such cases, the A-worlds closest to *w* are metaphysically impossible worlds). *Proposal 2* adds the following act-state-indepedence requirement to ECV-Max: in any decision situation, the truth-values of all option-to-outcome counterfactuals must be counterfactually independent of every option. S&W argue that Law-Bet$_1$ fails to meet this condition and conclude that it does not involve a "genuine decision" (24) and is therefore outside the scope of decision theory. They endorse an error theory to explain away the appearance that Law-Bet$_1$ is a real decision problem.

Unlike CERV-Max, SWT$_1$ gets Law-Bet$_1$ wrong. Revising ECV-Max in accordance with Proposal 1 or 2 yields an account that, unlike CERV-Max, gets Past-Bet$_1$ wrong. As S&W point out, to obtain an account that gets both cases right, we need to start with *SWT$_1$* (not with ECV-Max) and then revise the account by adopting either Proposal 1 or Proposal 2, thereby combining the revisions of ECV-Max intended to accommodate Past-Bet$_1$ and Law-Bet$_1$. Let SWT$_2$ (SWT$_3$) be the view resulting from revising SWT$_1$ by adopting Proposal 1 (Proposal 2).

SWT$_2$ and SWT$_3$ get both Past-Bet$_1$ and Law-Bet$_1$ right. However, unlike CERV-Max, they do not provide a *unified* solution to the problems the two cases pose for CDT. (The revisions of ECV-Max needed to get Past-Bet$_1$ right, i.e. those that lead to SWT$_1$, are quite different from those required to accommodate Law-Bet$_1$, i.e. Proposal 1 or 2.) What is worse, SWT$_2$ and SWT$_3$ (as well as ECV-Max and SWT$_1$) yield the wrong verdict in variants of Past-Bet$_1$ like the following.

> *Past-Bet$_3$*. A coin biased towards Tails was tossed yesterday in your absence. Arm amounts to accepting a bet on $P_1 \vee$ Heads, ¬Arm to accepting a bet on P_1 & Tails.
>
> P_1: The state of the universe yesterday at noon, combined with the natural laws of the actual world, metaphysically necessitates ¬Arm.
>
> You get \$1 for winning and \$0 for losing either bet. You know with certainty that determinism is true (so that Arm $\leftrightarrow$ ♦Arm and ¬Arm $\leftrightarrow$ ♦¬Arm are true) and that Heads & Arm, Heads & ¬Arm, Tails & Arm, and Tails & ¬Arm are nomically possible. Hence, you are also certain that each option is nomically compossible with winning and with losing. You have the following credences: Cr(Heads|P_1) = Cr(Heads|¬P_1) = Cr(Heads|Arm) = Cr(Heads|♦Arm) = Cr(Heads|¬Arm) = Cr(Heads |♦¬Arm) = Cr(Heads) = .4. Finally, you are certain that the coin-toss outcome is causally unaffected by your decision and counterfactually independent of Arm.

You are certain that Arm $\square\!\rightarrow$ \$1 holds iff Arm $\square\!\rightarrow$ ($P_1 \vee$ Heads) holds. Moreover, since P_1 and Heads are about yesterday, you are also certain that Arm $\square\!\rightarrow$ ($P_1 \vee$ Heads) holds iff $P_1 \vee$ Heads is true. Hence, by ECV-Max:

$$\begin{aligned} U(\text{Arm}) &= Cr(\text{Arm} \mathrel{\square\!\!\rightarrow} \$1) \\ &= Cr(P_1 \vee \text{Heads}) \end{aligned}$$

$$
\begin{aligned}
&= \mathrm{Cr}(P_1) + \mathrm{Cr}(\neg P_1 \,\&\, \mathrm{Heads}) \\
&= \mathrm{Cr}(P_1) + \mathrm{Cr}(\mathrm{Heads} \mid \neg P_1)\, \mathrm{Cr}(\neg P_1) \\
&= \mathrm{Cr}(P_1) + .4\ (1 - \mathrm{Cr}(P_1)) \\
&= .6\, \mathrm{Cr}(P_1) + .4
\end{aligned}
$$

By analogous reasoning:

$$
\begin{aligned}
\mathrm{U}(\neg \mathrm{Arm}) &= \mathrm{Cr}(\neg \mathrm{Arm} \mathbin{\square\!\!\rightarrow} \$1) \\
&= \mathrm{Cr}(P_1 \,\&\, \mathrm{Tails}) \\
&= \mathrm{Cr}(\mathrm{Tails} \mid P_1)\ \mathrm{Cr}(P_1) \\
&= .6\ \mathrm{Cr}(P_1)
\end{aligned}
$$

ECV-Max therefore entails that $\mathrm{U}(\mathrm{Arm}) > \mathrm{U}(\neg \mathrm{Arm})$. By SWT_1's lights, both outcomes (\$1 and \$0) are relevant to the evaluation of either option. SWT_1 therefore agrees with ECV-Max that $\mathrm{U}(\mathrm{Arm}) > \mathrm{U}(\neg \mathrm{Arm})$. Moreover, the argument from SWT_1 to the unique rationality of Arm that I just gave does not require the assumption that the closest Arm-worlds or the closest ¬Arm-worlds are metaphysically possible, nor does it require the assumption that these worlds feature violations of the actual laws. Therefore, SWT_1ists who revise their positions by endorsing Proposal 1 do not thereby gain any way of resisting the argument. So, SWT_2 also predicts that $\mathrm{U}(\mathrm{Arm}) > \mathrm{U}(\neg \mathrm{Arm})$. Finally, the truth-values of the four option-to-outcome counterfactuals are counterfactually independent of the options. Consider $\mathrm{Arm} \mathbin{\square\!\!\rightarrow} \$1$. We already saw that this counterfactual has the same truth-value as $P_1 \vee \mathrm{Heads}$. Note that this is true not just at the actual world, but also at the closest Arm-worlds and the closest ¬Arm-worlds. Consequently, since $P_1 \vee \mathrm{Heads}$ has the same truth-value at all these worlds, so does $\mathrm{Arm} \mathbin{\square\!\!\rightarrow} \$1$. $\mathrm{Arm} \mathbin{\square\!\!\rightarrow} \$1$ is therefore counterfactually independent of Arm and of ¬Arm. Similar reasoning applies to the three other option-to-outcome counterfactuals. This shows that SWT_1ists' judgments about Past-Bet_3 will not change if they revise their accounts by adopting Proposal 2. That is to say, SWT_3 also predict that $\mathrm{U}(\mathrm{Arm}) > \mathrm{U}(\neg \mathrm{Arm})$.

That prediction seems wrong. Being certain that Arm is nomically incompossible with P_1, you should ignore Arm-worlds where you win because P_1 holds. The only Arm & Win-worlds worth considering are Heads-worlds. Similarly, the ¬Arm & Lose-worlds to consider are Heads-worlds, not $\neg P_1$-worlds. Consequently, since the coin is biased towards Tails, $\mathrm{Cr}^{\mathrm{Arm}}(\mathrm{Win})$ and $\mathrm{Cr}^{\neg \mathrm{Arm}}(\mathrm{Lose})$ should be lower than $\mathrm{Cr}^{\mathrm{Arm}}(\mathrm{Lose})$ and $\mathrm{Cr}^{\neg \mathrm{Arm}}(\mathrm{Win})$, so that $\mathrm{U}(\neg \mathrm{Arm}) > \mathrm{U}(\mathrm{Arm})$. CERV-Max yields the correct result. Since you are certain of determinism, you are certain that P_1 is true iff ¬Arm holds iff $\blacklozenge \neg \mathrm{Arm}$ holds, so that $\mathrm{Cr}(P_1 \mid \blacklozenge \neg \mathrm{Arm}) = 1$. Moreover, $\mathrm{Cr}(P_1 \mid \blacklozenge \mathrm{Arm}) = 0$. Hence:

$$
\begin{aligned}
&\mathrm{CERV}(\mathrm{Arm}) = \mathrm{Cr}(\mathrm{Arm} \mathbin{\square\!\!\rightarrow} \$1 \mid \blacklozenge \mathrm{Arm}) = \mathrm{Cr}(P_1 \vee \mathrm{Heads} \mid \blacklozenge \mathrm{Arm}) = \mathrm{Cr}(\mathrm{Heads} \mid \blacklozenge \mathrm{Arm}) = .4 \\
&\mathrm{CERV}(\neg \mathrm{Arm}) = \mathrm{Cr}(\neg \mathrm{Arm} \mathbin{\square\!\!\rightarrow} \$1 \mid \blacklozenge \neg \mathrm{Arm}) = \mathrm{Cr}(P_1 \,\&\, \mathrm{Tails} \mid \blacklozenge \neg \mathrm{Arm}) = \mathrm{Cr}(\mathrm{Tails} \mid \blacklozenge \neg \mathrm{Arm}) = .6
\end{aligned}
$$

Joyce. James Joyce (2016) defends ECV-Max's predictions about a variant of Past-Bet_1:

Past-Bet$_4$. You are highly confident that the deterministic principle L is a law. As shown in Table 3, Arm amounts to accepting a bet on some proposition P_3 about yesterday with payoffs of \$10 and –\$1, and ¬Arm to accepting a bet on P_3 with payoffs of \$1 and –\$10. P_3 & L is metaphysically incompossible with Arm, and ¬P_3 & L with ¬Arm (as indicated by ⊥). You are certain that, if ~L holds, then the past and the laws are compossible both with Arm and with ~Arm.

	P_3 & L	¬P_3 & L	P_3 & ¬L	¬P_3 & ¬L
Arm	⊥	−\$1	\$10	−\$1
¬Arm	\$1	⊥	\$1	−\$10

Table 3. Past-Bet$_4$.

L is deterministic. Therefore, if L holds, then only one option is compossible with the unaffected facts. Joyce claims that you are then not really facing a choice. Your deliberations should therefore focus exclusively on evaluating the options conditional on ¬L. So, $U(\text{Arm}) = U(\text{Arm} \mid \neg L) = 10\ Cr(P_3 \mid \neg L) - Cr(\neg P_3 \mid \neg L) > Cr(P_3 \mid \neg L) - 10\ Cr(\neg P_3 \mid \neg L) = U(\neg\text{Arm} \mid \neg L) = U(\neg\text{Arm})$. Arm is uniquely rational.

I find it unobvious that you are not facing a real decision if the unaffected facts determine your action. (Couldn't they determine a specific action by determining that you make a genuine decision to perform that action?) Moreover, Joyce's claim that you should choose Arm in Past-Bet$_4$ seems wrong to me. The reasons for choosing ¬Arm in Past-Bet$_1$ appear to me to carry over to Past-Bet$_4$. In addition, it is not completely clear to me how to generalize Joyce's treatment of Past-Bet$_4$ to examples like Past-Bet$_1$. (In Past-Bet$_1$ you are *certain* that the unaffected facts determine your action. Does that mean that there is no point in deliberating about what to do?) However, any plausible generalization will likely get (In)determinism-Bet wrong.

(In)determinism-Bet. You must decide between betting on determinism (Bet_D) and betting on indeterminism (Bet_I). You have a very high credence (smaller than 1) in determinism, both unconditionally and conditional on Bet_D, on Bet_I, on ♦Bet_D, and on ♦Bet_I. You are certain that, if indeterminism actually holds, then indeterminism's truth is counterfactually independent of Bet_D and Bet_I. Table 4 shows the payoffs.

	Determinism	Indeterminism
Bet_D	\$1,000,000	\$0
Bet_I	−\$1,000,000	\$1

Table 4. (In)determinism bet.

By Joyce's reasoning, you are not facing a genuine decision under determinism (since the unaffected facts determine your action). You should therefore assess your options conditional on indeterminism and choose Bet_I. To me, that seems wrong.[31]

[31] For comments, suggestions, and objections, I am very grateful to Arif Ahmed, David Builes, Julianne Chung, Adam Elga, Alan Hájek, James Joyce, Daniel Muñoz, Daniel Nolan, Toby Solomon, Jack Spencer, Wolfgang Spohn, Timothy Williamson, the participants of a graduate seminar on decision theory I taught at Princeton in the spring of 2021, and the audiences at talks based on earlier versions of this paper that I gave between 2015 and 2021 at the joint conference of Humboldt Universität zu Berlin and Princeton University on Causation and Cognition, the Faculty Lunchtime Colloquium and the Humanities Council at Princeton University, the Belgrade Conference on Conditionals, the Hebrew University, the University of Louisville, the Conference of the German Research Foundation (DFG) (Research Unit "What if?") at the University of Konstanz, the Australasian Association of Philosophy Conference, the University of Melbourne, Monash University, Lingnan University, and at DEX VII at UC Davis.

References

Adams, E. (1975). *The Logic of Conditionals*. Dordrecht: Reidel. doi: 10.1007/978-94-015-7622-2

Ahmed, A. (2013a). Causal Decision Theory and the Fixity of the Past. *British Journal for the Philosophy of Science*, 6, 665−685. doi: 10.1093/bjps/axt021

Ahmed, A. (2013b). Causal Decision Theory: A Counterexample. *Philosophical Review*, 122, 289−306. doi: 10.1215/00318108-1963725

Ahmed, A. (2014a). Dicing with Death. *Analysis*, 74, 587−592. doi: 10.1093/analys/anu084

Ahmed, A. (2014b). *Evidence, Decision and Causality.* Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139107990

Ahmed, A. (2015). Infallibility in the Newcomb Problem. *Erkenntnis*, 80, 261−273. doi: 10.1007/s10670-014-9625-x

Ahmed, A. (2021). *Evidential Decision Theory*. Cambridge: Cambridge University Press. doi: 10.1017/9781108581462

Ahmed, A. & Spencer, J. (2020). Objective Value is Always Newcombizable. *Mind*, 129, 1157−1192. doi: 10.1093/mind/fzz070

Albert, D. (2015). *After Physics*. Cambridge, MA: Harvard University Press. doi: 10.4159/harvard.9780674735507

Audi, P. (2012). A clarification and defence of the notion of grounding. In F. Correia & B. Schnieder (Eds.), *Metaphysical Grounding: Understanding the Structure of Reality* (pp. 101–21). Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139149136.004

Bassett, R. (2015). A critique of benchmark theory. *Synthese*, 192, 241–267. doi: 10.1007/s11229-014-0566-3

Bennett, J. (1984). Counterfactuals and Temporal Direction. *Philosophical Review*, 93, 57–91. doi: 10.2307/2184413

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Clarendon. doi: 10.1093/0199258872.003.0001

Briggs, R. (2012). Interventionist Counterfactuals. *Philosophical Studies*, 160, 139−166. doi: 10.1007/s11098-012-9908-5

Collins, J., Hall, N. & Paul, L.A. (eds.) (2004). *Causation and Counterfactuals*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1752.003.0001

Dorr, C. (2016). Against Counterfactual Miracles. *Philosophical Review*, 125, 241−286. doi: 10.1215/00318108-3453187

Edgington, D. (1995). On Conditionals. *Mind*, 104, 235−329. doi: 10.1007/978-1-4020-6324-4_3

Edgington, D. (2003). Counterfactuals and the Benefit of Hindsight. In P. Dowe & P. Noordhof (Eds.), *Causation and Counterfactuals* (pp. 12−27). London: Routledge.

Edgington, D. (2011). Conditionals, Causation, and Decision. *Analytic Philosophy*, 52, 75−87. doi: 10.1111/j.2153-960x.2011.00520.x

Egan, A. (2007). Some Counterexamples to Causal Decision Theory. *Philosophical Review*, 116, 93−114. doi: 10.1215/00318108-2006-023

Elga, A. (2022). Confession of a Causal Decision Theorist. *Analysis*, 82, 203−213. doi: 10.1093/analys/anab040

Fine, K. (2012). Guide to ground. In F. Correia & B. Schnieder (Eds.), *Metaphysical Grounding: Understanding the Structure of Reality* (pp. 37–80). Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139149136.002

Gibbard, A. & Harper, W. (1978). Counterfactuals and Two Kinds of Expected Utility. In A. Hooker, J. J. Leach & E. F. McClennen (Eds.), *Foundations and Applications of Decision Theory* (pp. 125−162). Dordrecht: Reidel. doi: 10.1017/cbo9780511609220.022

Goodman, J. (2015). Knowledge, Counterfactuals, and Determinism. *Philosophical Studies*, 172, 2275−2278. doi: 10.1007/s11098-014-0409-6

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and Counterfactuals* (pp. 225−276). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1752.003.0010

Hall, N. (2007). Structural Equations and Causation. *Philosophical Studies*, 132, 109–136. doi: 10.1007/s11098-006-9057-9

Hitchcock, C. (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *Journal of Philosophy*, 98, 273–299. doi: 10.2307/2678432

Hitchcock, C. (2013). What is the 'Cause' in Causal Decision Theory?. *Erkenntnis*, 78, 129−146. doi: 10.1007/s10670-013-9440-9

Hiddleston, E. (2005). A Causal Theory of Counterfactuals. *Noûs*, 39, 632−657. doi: 10.1111/j.0029-4624.2005.00542.x

Jackson, F. (1991). Decision-Theoretic Consequentialism and the Nearest and Dearest Objection. *Ethics*, 101, 461−482. doi: 10.4324/9780203019467-23

Joyce, J. (1999). *The Foundations of Causal Decision Theory*, Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511498497

Joyce, J. (2016). Review of Arif Ahmed, Evidence, Decision and Causality. *Journal of Philosophy*, 113, 224−232. doi: 10.5840/jphil2016113413

Kment, B. (2006). Counterfactuals and Explanation. *Mind*, 115, 261−310. doi: 10.1093/mind/fzl261

Kment, B. (2010). Causation: Determination and Difference-Making. *Noûs*, 44, 80−111. doi: 10.1111/j.1468-0068.2009.00732.x

Kment, B. (2014). *Modality and Explanatory Reasoning*, Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199604685.001.0001

Koslicki, K. (2012). Essence, necessity and explanation. In T. Tahko (Eds.), *Contemporary Aristotelian Metaphysics* (pp. 187–206). Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511732256.014

Levi, I. (1975). Newcomb's many problems. *Theory and Decision*, 6, 161−175. doi: 10.1007/978-94-015-1121-6_15

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, D. (1976). Probabilities of conditionals and conditional probabilities. *Philosophical Review*, 85, 297−315. doi: 10.1007/978-94-009-9117-0_6

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Noûs*, 13, 455−476. doi: 10.2307/2215339

Lewis, D. (1981). Causal Decision Theory. *Australasian Journal of Philosophy*, 59, 5−30. doi: 10.1080/00048408112340011

Lewis, D. (1986). Postscripts to "Causation". In *Philosophical Papers*, vol. II (pp. 172–213). Oxford: Oxford University Press.

Lewis, D. (2004). Causation as Influence. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and Counterfactuals* (pp. 75–107). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1752.003.0004

Loewer, B. (2007). Counterfactuals and the Second Law. In H. Price & R. Corry (Eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (pp. 293–326). Oxford: Oxford University Press.

McDermott, M. (1995). Redundant Causation. *British Journal for the Philosophy of Science*, 46, 523–544. doi: 10.1093/bjps/46.4.523

Nolan, D. (2017). Causal Counterfactuals and Impossible Worlds. In H. Beebee, C. Hitchcock & H. Price (Eds.), *Making a Difference* (pp. 14–32). Oxford: Oxford University Press. doi: 10.1093/oso/9780198746911.003.0002

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (pp. 114−146). Dodrecht: Reidel. doi: 10.1007/978-94-017-1466-2_7

Nute, D. (1980). *Topics in Conditional Logic*. Dordrecht: Springer. doi: 10.1007/978-94-009-8966-5

Paul, L. A. (2004). Aspect Causation. In J. Collins, N. Hall & L.A. Paul (Eds.), *Causation and Counterfactuals* (pp. 205–224). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1752.003.0009

Paul, L. A. & Hall, N. (2013). *Causation: A User's Guide*. Oxford: Oxford University Press.

Railton, P. (1986). Alienation, Consequentialism, and the Demands of Morality. *Philosophy & Public Affairs*, 13, 134−171. doi: 10.4324/9780203723746-38

Rosen, G. (2010). Metaphysical dependence: grounding and reduction. In B. Hale & A. Hoffmann (Eds.), *Modality: Metaphysics, Logic, and Epistemology* (pp. 109–136). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199565818.003.0007

Sandgren, A. & Williamson, T. (2020). Determinism, Counterfactuals, and Decision. *Australasian Journal of Philosophy*, 99, 286−302. doi: 10.1080/00048402.2020.1764073

Schaffer, J. (2004). Counterfactuals, Causal Independence and Conceptual Circularity. *Analysis*, 64, 299−309. doi: 10.1093/analys/64.4.299

Schaffer, J. (2009). On what grounds what. In D. Chalmers, D. Manley & R. Wasserman (Eds.), *Metametaphysics: New Essays on the Foundations of Ontology* (pp. 347–383). Oxford: Clarendon Press. doi: 10.1093/pq/pqx031

Seidenfeld, T. (1984). Comments on Causal Decision Theory. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2, 201–212. doi: 10.1086/psaprocbienmeetp.1984.2.192505

Skyrms, B. (1980). *Causal Necessity*. Yale University Press. doi: 10.2307/2215225

Slote, M. (1978). Time in Counterfactuals. *Philosophical Review*, 87, 3–27. doi: 10.2307/2184345

Sobel, J. 1988. Infallible Predictors. *Philosophical Review*, 97, 3–24. doi: 10.2307/2185097

Solomon, T. (2019). Causal Decision Theory's Pre-Determination Problem. *Synthese*, 198, 5623−5654. doi: 10.1007/s11229-019-02425-0

Spencer, J. & Wells, I. (2019). Why Take Both Boxes?. *Philosophy and Phenomenological Research*, 99, 27−48. doi: 10.1111/phpr.12466

Spencer, J. (2021a). Rational Monism and Rational Pluralism. *Philosophical Studies*, 178, 1769−1800. doi: 10.1007/s11098-020-01509-9

Spencer, J. (2021b). An Argument Against Causal Decision Theory. *Analysis*, 81, 52−61. doi: 10.1093/analys/anaa037

Stalnaker, R. (1968). A Theory of Conditionals. In N. Rescher (Ed.), *Studies in Logical Theory* (pp. 98−112). American Philosophical Quarterly Monograph Series 2. Oxford: Blackwell. doi: 10.1093/oso/9780198810346.003.0010

Stalnaker, R. (1978). Letter to David Lewis, May 21, 1972. In A. Hooker, J. J. Leach & E.F. McClennen (Eds.), *Foundations and Applications of Decision Theory* (pp. 151−152). Dordrecht: Reidel. doi: 10.1007/978-94-009-9117-0_7

Tichý, P. (1976). A Counterexample to the Stalnaker-Lewis Analysis of Counterfactuals. *Philosophical Studies*, 29, 271–273. doi: 10.1007/bf00411887

Wasserman, R. (2006). The Future Similarity Objection Revisited. *Synthese*, 150, 57−67. doi: 10.1007/s11229-004-6256-9

Wedgwood, R. (2013). Gandalf's Solution to the Newcomb Problem. *Synthese*, 190, 2643−2675. doi: 10.1007/s11229-011-9900-1

Williamson, T. & Sandgren, A. 2021. Law-Abiding Causal Decision Theory. *British Journal for the Philosophy of Science*. doi: 10.1086/715103

Wilson, J. (2014). Hume's Dictum and the Asymmetry of Counterfactual Dependence. In A. Wilson, ed., *Chance and Temporal Asymmetry* (pp. 258–279). Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780199673421.003.0013