Counterfactuals and the Gibbard-Harper Collapse Lemma

Melissa Fusco
Columbia University

September 17, 2019

Gibbard & Harper (1978) provides a classic statement of Causal Decision Theory ("CDT"), which uses counterfactual conditionals to express the causal relationships that are, according to CDT, of particular relevance to rational decision-making. The account builds a bridge between decision theory and the semantics of natural language counterfactuals, active at the time in the work of Lewis and Stalnaker and still vibrant today.[1]

CDT's rival in the dialectic in which Gibbard & Harper are situated is Evidential Decision Theory ("EDT"). While EDT, like CDT, holds that any choiceworthy act is one which maximizes expected utility, EDT employs act-conditionalized probabilities in the calculation of expected utility. This is typically conceived of as an attitude of austerity towards the causation-correlation distinction: while it may be perfectly real, it has no important direct role to play in a theory of decision.

Classic CDT — in particular, the Gibbard-Harper formulation of it — has enjoyed wide acceptance. Many in the recent literature, however, hold that tides are turning. One factor in the sea-change is an influential 2007 paper by Andy Egan (Egan, 2007), which presents several counterexamples to the theory. On Egan's telling, causal decision theorists — and he does have in mind those who appeal to the counterfactual formulation of the theory[2] — adhere to the motto "do whatever has the best expected outcome, holding fixed your initial views about the likely causal structure of the world" (96). However, Egan argues, there are cases where agents should *not* hold such initial views fixed as they act. In such cases, agents should use their anticipated future causal views instead, taking into account what they expect to *learn* by performing the very act in question.

In this paper, I focus on the dialectic from the CDTer's point of view, with an eye to a formal result pointed out by Gibbard & Harper in the third section of their classic paper. There, they show that if an agent's credences are probabilistically coherent, and the semantics for counterfactuals obeys Strong Centering — roughly, the view that each

[1] Gibbard & Harper cite, in particular, Lewis (1973), Stalnaker (1968). For recent work in this tradition, see, *inter alia*, Ahmed (2013); Kment (2019).

[2] See Egan (2007, pg. 95).

possible world is counterfactually closest to itself — then the probability of (a counterfactual conditional on its antecedent) simplifies to the probability of (its consequent, given its antecedent). This has the eyebrow-raising consequence that "Eganized" causal decision theory, the view on which agents anticipate their future causal views, recommends an act just in case classical *evidential* decision theory does.

The "collapse", as I call it, complicates the traditional way of glossing the relationship between EDT, CDT, and diachronic coherence norms. I canvas three takes on Gibbard & Harper's discussion of the result in §4 below, arguing in favor of one which emphasizes the peculiarity of predicaments involving choosing one's own evidence. This peculiarity raises doubts about whether update in Egan-type cases can aspire to the status of probabilistic *knowledge*, in the sense of Moss (2013a, 2018). I suggest that they do not, and explain why this consideration both points the way to understanding the true significance of the collapse, and functions as a defense of CDT against Egan's counterexamples.

1 EDT vs. CDT: an overview

1.1 Decision Problems

Both classical EDT and classical CDT begin with the thought that the value of each of a set of available acts (call them the a's) can be calculated by identifying a set of states which fix one's welfare (call them the s's) and then multiplying the utility of each state-act conjunction by one's subjective probability, or *credence*, that that state obtains.

For example, suppose that Otto's tennis match is today. Calliope is offered a bet on his winning at even odds for a dollar: she can either *bet on Otto* (henceforth B) or decline the bet (= $\neg B$). The payoff of the bet depends on whether *Otto wins* (= W) or not (= $\neg W$). In matrix form, her possible payoffs look like this:

	Otto wins (W)	Otto loses ($\neg W$)
Bet (B)	\$1	-\$1
Don't bet ($\neg B$)	\$0	\$0

According to both decision theories, Calliope is facing a *decision problem* in which her goal is *maximize expected utility*. Her noninstrumental desires equip her with a value function $Val(\cdot)$ over states called *outcomes*, which — following standard idealization — I will assume does not change over time, and is such that $Val(\$nk) = kVal(\$n)$. Her decision problem at a time t can be represented as a triple $\langle Cr^t, \mathcal{A}, \mathcal{S}\rangle$, where

> $Cr^t(\cdot)$ is a credence function over the state space W, here representing Calliope's subjective confidence in a variety of propositions at t;
>
> $\mathcal{A} = \{a_1, \ldots, a_n\}$ is a partition of W into Calliope's available *acts* at t, and
>
> $\mathcal{S} = \{s_1, \ldots, s_m\}$ is a partition of W into admissible *states of nature*, where a partition is admissible only if each act-state conjunction $(a_i \wedge s_j)$ determines

some number $Val(a_i \wedge s_j)$ (unique up to positive affine transformation) under the value function.[3]

Epistemologists of varying stripes will take the first element of the decision problem, the agent's credence function $Cr^t(\cdot)$, to be subject to a variety of *epistemic norms*. Of particular note are, first, **Probabilism**: $Cr^t(\cdot)$ should be a probability function. Second, $Cr(\cdot)$ is commonly taken to be governed by **Conditionalization**, a diachronic norm which concerns how the agent should respond to new information. Conditionalization states that an agent who, between times t and t^+, learns exactly E, should adopt the posterior credence function $Cr^{t^+}(\cdot) = Cr^t(\cdot \mid E)$, where this is defined.[4]

For our purposes, a more general norm is also worth mentioning, which extends Conditionalization to cases where a learning experience does not result in an agent's becoming *certain* of any proposition E. **Jeffrey Conditionalization** states that an agent who, between times t and t^+, undergoes a learning experience which directly alters her credences in members of the partition $\{E_i\}$ from $Cr^t(E_i)$ to $Cr^{t+}(E_i)$, should adopt the posterior credence function $Cr^{t+}(\cdot) = \sum_i Cr^{t+}(E_i)Cr^t(\cdot \mid E_i)$. (Jeffrey Conditionalization will become relevant in §4.2 below.) There may be—and discussions in the literature often presuppose that there are—further, less purely subjective norms governing credence functions, such as that they are based on reasonable priors, sufficiently sensitive to the observed frequencies of events, and so on.

1.2 The EDT Branch

EDT and CDT's common starting point is Savage (1972)'s notion of expected utility, which is simply the generic statistical notion of expected value, applied to the value function.[5] His theory says: *in an uncertain world, estimate the value of act a_j by taking the sum of its value in each state of nature, weighted by one's current estimate that that state obtains.*

Equation 1 (Savage Expected Utility). $SEU^t(a_j) = \sum_i Cr^t(s_i)Val(a_j \wedge s_i)$.

Savage's decision rule: maximize expected utility at t by choosing an act $a_j \in \mathcal{A}$ such that $SEU^t(a)$ is maximal.[6]

Savage's theory entails the validity of *dominance arguments* in favor of a particular acts. An act a_j dominates all other acts $a_i \in \mathcal{A}$ when, for all states s, $Val(s \wedge a_j) > Val(s \wedge a_i)$. The payoff of a_j is thus greater than the payoff of *any* other a_i in *any* state with positive probability, and Savage's theory will recommend a_j.

[3] In what follows, I will speak of the members of $\mathcal{A}$, $\mathcal{S}$, and the domain of $Cr^t(\cdot)$ alike as *propositions*. This differs from Savage's original picture of the relevant primitives, on which acts are *functions* from states to outcomes. See Joyce (1999, Ch. 2) for discussion of this shift.

[4] Conditional probabilities of the form $Pr(A \mid B)$ are customarily defined, via the "Ratio Formula", to be $Pr(A \wedge B)/Pr(B)$. On primitive treatments of conditional probability, such as Spohn (1986), the Ratio Formula equality does not *define* conditional probability, but holds whenever $Pr(B) \neq 0$.

[5] Where F is a function, the generic notion of expected value says that $E[F] = \sum_i f_i Pr(F = f_i)$. Here, our probability function is the subjective credence function $Cr^t(\cdot)$, and the function F is $Val(\cdot)$ across act-state pairs, where the act is held fixed.

[6] That is, such that $SEU^t(a_j) \geq SEU^t(a_i)$, for any $a_i \in \mathcal{A}$.

While granting that Savage's norm works for some decision problems, both CDT and EDT move away from it as a general decision rule. One way to see why is to observe that there are cases in which the agent believes the likelihood of a state W (Otto's winning) *depends* on whether she performs an act like B (taking the bet).

Recall that, in Calliope's case, she should take the bet on Otto at t just in case the expected utility of B exceeds the expected utility of $\neg B$. According to Savage's theory, this happens just in case $SEU^t(B) > SEU^t(\neg B)$, which happens in this instance just in case $Cr^t(W) > Cr^t(\neg W)$.[7] Suppose Calliope knows that Otto's confidence increases whenever he sees her betting on him: based on her data, he has a 70% chance of winning if she accepts B, but only a 40% chance otherwise. Assuming Probabilism, Calliope's current best estimate of Otto's likelihood of winning, $Cr^t(W)$, is the weighted average of the probability that he wins, *given that she bets on him*, and the probability that he wins, *given that she doesn't*, where the "weights" are her unconditional credences in her own acts B and $\neg B$, respectively:

$$\begin{aligned} Cr^t(W) &= Cr^t(W \wedge B) + Cr^t(W \wedge \neg B) \\ &= Cr^t(W \mid B)\underline{Cr^t(B)} + Cr^t(W \mid \neg B)\, Cr^t(\neg B) \\ &= .7 \times \underline{Cr^t(B)} + .4 \times \; Cr^t(\neg B) \end{aligned}$$

But Calliope doesn't usually place bets, so her initial credence in B is only 20%. This attaches a small weight to the highish probability she assigns to Otto's winning conditional on her bet—and a *large* weight to the lowish probability she assigns to Otto's winning conditional on her *not* betting. A flatfooted application of Savage's theory will therefore recommend against B. This seems obviously wrong. In using Savage's equation to assign expected utility to B, Calliope is improperly diluting the probability assigned to Otto's winning by including in her calculations worlds where she doesn't take the bet. Calliope wants to know what the expected utility of *the act of betting on Otto* is; in that case, though, it is certainly true — with probability 1, not probability .2 — that B occurs.

Savage's own response to this problem was to require that decision problems be formulated with a special partition $\mathcal{S}^*$ of states that are known to be *independent* of the agent's contemplated acts.[8] But it isn't always clear that such an $\mathcal{S}^*$ can be found. The

[7] Calculation:

$$\begin{aligned} & SEU^t(B) > SEU^t(\neg B) \\ \text{iff } & SEU^t(B) > 0 \text{ (no money changes hands if Calliope declines to bet)} \\ \text{iff } & \sum_i Cr^t(s_i) Val(B \wedge s_i) > 0 \\ \text{iff } & Cr^t(W) \times 1 + Cr^t(\neg W) \times -1 > 0 \\ \text{iff } & Cr^t(W) > Cr^t(\neg W). \end{aligned}$$

[8] Is the relevant type of independence *causal*, or *evidential*? At stake is, once again, the difference between EDT and CDT. My understanding is that there is controversy over which form of independence Savage intended (see e.g. Jeffrey (1983, pgs. 21-22) and Joyce (1999, §4.1)).

evidential decision theorist takes a different path: when assigning expected utility to an option a, use any partition $\mathcal{S}$ you like, but do not use your current credence in $s_i \in \mathcal{S}$ — rather, use your credence in s_i, *conditional on* a.

Equation 2 (Evidential Expected Utility). $\mathcal{V}^t(a_j) = \sum_i Cr^t(s_i \mid a_j) Val(a_j \wedge s_i)$.
Conditional decision rule: maximize expected utility by choosing an act $a \in \mathcal{A}$ such that $\mathcal{V}^t(a)$ is maximal.

With respect to the example above, this change removes the problematic weighting by Calliope's (low) current credence in B. The factor by which $Cr^t(W \mid B)$ is now weighted is simply 1. Given this way of calculating expected utility, the result for B is positive, so EDT recommends that Calliope take the bet.

1.3 The CDT Branch

In the case of Calliope and Otto, it is natural to assume that B (Calliope's bet) and W (Otto's winning) are probabilistically correlated because Calliope's bet on Otto *causally conduces* to his winning (perhaps by increasing his confidence). But not all correlation is causation. CDT diverges from EDT by insisting that causal information be represented *separately* from evidential support, and appealing to act-conditioned probabilities — applying a different probability function to different acts — *only* when evidential support is also causal.

Here is the sort of case where the two theories diverge. Suppose Adeimantus is hoping to get an REI jacket from his mother for Christmas (=J). He begins to contemplate *leaving REI catalogues around the house* (=C), on the grounds that, statistically, houses full of REI catalogues are more likely to have REI jackets inside. If buying catalogues and planting them in the house costs 1 utile, Adeimantus's outcomes look like this:

	Jacket (J)	No jacket ($\neg J$)
Catalogues (C)	9	0
No catalogues ($\neg C$)	10	1

$$Cr^t(J \mid C) > Cr^t(J)$$

A calculation by the conditional decision norm will recommend that Adeimantus see to it that there are catalogues around the house, so long as C raises the statistical probability of J by more than 1/9.[9]

Suppose, however, that Adeimantus believes his mother has *already* purchased the gift at time t. In this case, it isn't clear that EDT makes the right recommendation. The

[9] Calculation:

$$\begin{gathered}
\mathcal{V}^t(C) > \mathcal{V}^t(\neg C) \text{ iff} \\
\sum_i Cr^t(k_i \mid C) Val(k_i \wedge C) > \sum_i Cr^t(k_i \mid \neg C) Val(k_i \wedge \neg C) \text{ iff} \\
[Cr^t(J \mid C) \times 9 + Cr^t(\neg J \mid C) \times 0] > [Cr^t(J \mid \neg C) \times 10 + Cr^t(\neg J \mid \neg C) \times 1] \text{ iff} \\
[Cr^t(J \mid C) \times 9] > [Cr^t(J \mid \neg C) \times 10 + Cr^t(\neg J \mid \neg C)]
\end{gathered}$$

causal decision theorist will grant that, since $Cr^t(J \mid C) > Cr^t(J)$, it would be good for Adeimantus to spontaneously *discover*, or *receive the news*, that there are REI catalogues around the house. That is why the CDTer calls the EDTer's decision-making quantity, $\mathcal{V}^t(\cdot)$, a "news value" function.[10] The problem, the CDTer says, is that *receiving the news* that there are REI catalogues around the house should be treated differently than *making it the case* that the very same thing obtains. What matters for a decision problem is how likely an available act a is to *cause* some state — such as J — that the agent desires. And as Adeimantus himself believes, it cannot cause that state, since all gifts have already been purchased.

In light of cases like this, Gibbard and Harper advance a different utility-maximizing equation, wherein the relevant subjective probability is $Cr(a_j \boxdot\!\!\rightarrow s_k)$. I will call the object of the agent's credence here the *act-counterfactual* $\ulcorner a_j \boxdot\!\!\rightarrow s_k \urcorner$, and follow Gibbard & Harper in reading it as the subjective probability that *if act a_i were performed, state s_k would obtain*:

Equation 3 (Causal Expected Utility)**.** $\mathcal{U}^t(a_j) = \sum_i Cr^t(a_j \boxdot\!\!\rightarrow s_i) Val(a_j \wedge s_i)$.
Causal decision rule: maximize expected utility by choosing an act $a \in \mathcal{A}$ such that $\mathcal{U}^t(a)$ is maximal.

In our example, Adeimantus's belief that leaving REI catalogues around the house at t has no causal influence over whether J obtains is reflected in the fact that his credences satisfy what we will call the *Counterfactual Independence Criterion*:

Definition 1 (Counterfactual Independence Criterion)**.** An agent believes, at t, that act a has no causal power over state s iff $Cr^t(s) = Cr^t(a \boxdot\!\!\rightarrow s) = Cr^t(\neg a \boxdot\!\!\rightarrow s)$.[11]

In the case at hand, Adeimantus's credences are such that $Cr^t(J) = Cr^t(C \boxdot\!\!\rightarrow J) = Cr^t(\neg C \boxdot\!\!\rightarrow J)$. Likewise, $Cr^t(\neg J) = Cr^t(C \boxdot\!\!\rightarrow \neg J) = Cr^t(\neg C \boxdot\!\!\rightarrow \neg J)$. The CDTer will thus accept a Savage-style dominance argument to the effect that, no matter

---

Since $Cr^t(\cdot \mid C)$ and $Cr^t(\cdot \mid \neg C)$ are themselves probability functions, this is equivalent to

$$[Cr^t(J \mid C) \times 9] > [Cr^t(J \mid \neg C) \times 10 + (1 - Cr^t(J \mid \neg C))]$$

Letting $n = Cr^t(J \mid C)$ and $m = Cr^t(J \mid \neg C)$, this obtains just in case

$$\begin{aligned} 9n &> 10m + (1 - m) \text{ iff} \\ n &> m + 1/9. \end{aligned}$$

[10] For use of the term "news value" to describe $\mathcal{V}(a)$, see e.g. Lewis (1981), Gibbard & Harper (1978), and Joyce (1999).

[11] See Gibbard & Harper (1978, pg. 136, paragraph 2). Note that for *conditional* probabilities, that $Cr^t(s) = Cr^t(s \mid a)$ entails $Cr^t(s) = Cr^t(s \mid \neg a)$. The analogous entailment holds for counterfactuals given Gibbard & Harper's Axiom 2 (op cit., pg. 128), but fails easily on Lewis (1973)'s more general treatment of the semantics of counterfactuals.

what Adeimantus's time-t credence in J is, he should refrain from planting catalogues in the house.[12] He should save himself the one-utile effort of bringing about C.

1.4 Newcomb Problems and NC Problems

We are now in a position to describe *NC Problems*, of interest to us because they are the simplest cases in which EDT and CDT disagree "on the ground." For our purposes, an NC problem will be a problem with two acts $\{a, \neg a\}$ and two states $\{s, \neg s\}$ for which the Counterfactual Independence Criterion holds with respect to credence function $Cr(\cdot)$ at time t. Nonetheless, the agent's relevant conditional credences — perhaps because she defers to reliability — are such that a is a very good indication of s: $Cr^t(s \mid a) \gg Cr^t(s) \gg Cr^t(s \mid \neg a)$. In such a situation, it is easy to engineer the stakes so that CDT and EDT come apart. It is sufficient, for example, to "sweeten" $\neg a$, making it slightly better $(+\Delta)$ than a both in s and in $\neg s$, while making state s considerably more valuable $(+\Theta)$ than $\neg s$, no matter whether a holds:

	s	$\neg s$
a	Θ	0
$\neg a$	$\Theta + \Delta$	Δ

Figure 1: NC Problem Payoffs. Shading indicates high conditional probability.

The high conditional probability of s, given a, gives the EDTer decisive reason to choose a in this problem, the lack of causal influence between a and s notwithstanding. On the other hand, the sweetener Δ and the causal independence of $\neg a$ from s gives the CDTer decisive reason to choose $\neg a$, the statistical support $\neg a$ lends to $\neg s$ notwithstanding.

The classic statement an NC Problem, of course, involves a wizardly predictor:

> **Newcomb's Puzzle**. There are two boxes before you, a large opaque box and a small clear box containing \$1,000. You may *take both boxes* ($= 2B$), or *take just the opaque box* $(= 1B)$, keeping whatever is inside the box(es) you take. But: an uncannily accurate predictor has put either \$1 million

[12] Calculation: by Equation 3 and the Counterfactual Independence Criterion,

$$\mathcal{U}^t(\neg C) > \mathcal{U}^t(C) \text{ iff}$$
$$\sum_i Cr^t(\neg C \mathrel{\square\!\!\rightarrow} s_i) Val(\neg C \wedge s_i) > \sum_i Cr^t(C \mathrel{\square\!\!\rightarrow} s_i) Val(C \wedge s_i) \text{ iff}$$
$$\sum_i Cr^t(s_i) Val(\neg C \wedge s_i) > \sum_i Cr^t(s_i) Val(C \wedge s_i).$$

But this is immediate, by the dominance structure of the payoff matrix where $\mathcal{S} = \{J, \neg J\}$ (repeated):

	J	$\neg J$
C	9	0
$\neg C$	10	1

or \$0 in the opaque box. She has put the million in the opaque box just in case *she predicted you would take one box* (= $P1$), and withheld the million just in case *she predicted you would take both boxes* (= $P2$). (Nozick, 1969; Gibbard & Harper, 1978, §10)

	$P1$	$P2$
$1B$	\$1 million	\$0 million
$2B$	\$1.001 million	\$0.001 million

$$Cr^t(P1 \mid 1B) \gg Cr^t(P1) \gg Cr^t(P1 \mid 2B)$$

2 Egan's Case, and its diachronic bite

While Egan argues that CDT is the wrong theory of decision, he concedes that it delivers the right verdict in Newcomb's puzzle (Egan, 2007, pg. 94): in light of the fact that the million dollars has already been distributed (or withheld), it is better to take both boxes. However, he disagrees with similar reasoning in other cases. His counterexample to Gibbard & Harper-style CDT goes as follows:[13]

> **Murder Lesion**. Mary is deliberating about whether to shoot the tyrant Alfred. She would prefer to shoot him, but only if she will hit him, rather than miss him. Mary has good evidence that a certain kind of brain lesion, which she may or may not have, causes murderous tendencies but also causes shooters to have bad aim. Mary currently has high credence that she has good aim. But (like Calliope in §1) she assigns low credence to the proposition that she will act.

	Don't hit ($\neg H$)	Hit (H)
Shoot (S)	terrible	good
Don't shoot ($\neg S$)	neutral	*(impossible)*

In the Murder Lesion case, the available acts are: shoot or don't ($\{S, \neg S\}$) and the basic states are: hit or don't ($\{H, \neg H\}$). However, Mary's knowledge includes information about causal influence: she has conditional and unconditional subjective probabilities on well-formed formulas like $\ulcorner S \boxdot\!\!\rightarrow H \urcorner$ — *if I were to shoot Alfred, I would hit him.* Egan suggests that the set of states used to calculate Mary's expected utility should be

$$\{S \boxdot\!\!\rightarrow \neg H, S \boxdot\!\!\rightarrow H\}$$

[13] This section, and the next, reproduce at slightly greater length the arguments given in Fusco (2017).

rather than $\{H, \neg H\}$. This gives rise to a second matrix:

	$S \boxrightarrow \neg H$	$S \boxrightarrow H$
S	terrible	good
$\neg S$	neutral	neutral

However, the second matrix has a striking feature: although Mary takes the causal relationships across the top to be causally independent of her acts, her credences in them will change drastically if she actually shoots. Egan argues on this basis that the act-conditional probability relevant to the calculation of expected utility is given by the more complex formula (*):

$$Cr((a_i \boxrightarrow s_k) \mid a_i). \qquad (*)$$

In (*), a_i appears twice, and both subjunctive and evidential probability are invoked. (*) should be read as: *the subjective probability that (e.g.) if Mary were to shoot Alfred, she would hit him, given that she shoots.* I will henceforth call (*) "the Egan credence on s_k in a_i." Calculating the expected utility of an act a_i with Egan credences in state-act pairs yields a quantity we can call $\mathcal{U}_{\text{Egan}}(a_i)$:

Equation 4. $\mathcal{U}^t_{\text{Egan}}(a_i) = \sum_k Cr^t(a_i \boxrightarrow s_k \mid a_i) Val(s_k \wedge a_i)$.
Eganized causal decision rule: maximize expected utility by choosing an act $a \in \mathcal{A}$ such that $\mathcal{U}^t_{\text{Egan}}(a)$ is maximal.

Egan argues that, intuitively, Mary should *not* shoot in *Murder Lesion*, and that the equation for $\mathcal{U}_{\text{Egan}}$ delivers this result. As Mary confronts her decision, $Cr(S)$, by description of the case, is low. Therefore, $Cr(S \boxrightarrow H)$ is high, since she is relatively confident she does not have the brain lesion. Finally, the Egan credence (*) = $Cr((S \boxrightarrow H) \mid S)$ is low, since once Mary conditionalizes on *shoot*, she is quite confident she has the lesion, and the lesion causes bad aim. Applying the Eganized Causal Credence norm, we get Egan's favored answer, which is that shooting has a low expected utility. By contrast, classical CDT uses the unconditional probability of $\ulcorner S \boxrightarrow H \urcorner$, which — wrongly, Egan says — predicts the expected utility of shooting to be high.[14]

Having presented these examples, it is worth briefly reviewing the dialectic. Egan *follows* Gibbard & Harper in holding that counterfactuals have a role to play in describing the subjective probability relevant to the expected value of an act a given a partition $\mathcal{S}$. Although he frames his case as a counterexample to classical CDT, the case involves an

[14] Calculation sketches (suppressing the subscript 't' on credence and utility functions):

$$\begin{aligned}\mathcal{U}(S) &= \sum_k Cr(s_k) Val(s_k \wedge S) \\ &= Cr(S \boxrightarrow H) Val(S \wedge H) + Cr(S \boxrightarrow \neg H) Val(S \wedge \neg H) \\ &= (\textbf{high} \times Val(H)) + (\textbf{low} \times Val(\neg H))\end{aligned}$$

agent who has credences concerning causal influence — credences that are appealed to in deliberation.

What Egan's considerations add, though, is the diachronic norm of Conditionalization, previously mentioned but hitherto not explicitly invoked. Conditionalization entails that an agent's conditional credences remain constant over the course of a learning experience on E, a feature known as *rigidity*:

Fact 1. (Conditional Rigidity). As an ideal agent undergoes a learning experience on E between any t and $t^+ \geq t$,

$$Cr^t(\phi \mid E) = Cr^{t^+}(\phi \mid E)$$

By contrast, and as Gibbard & Harper emphasize, $Cr(a \ \square\!\!\rightarrow \ s)$ — the CDTer's proxy for credence in the counterfactual, ⌜if I *were* to do a, then it *would* be the case that s⌝ — is *non-rigid*: it varies with $Cr(s)$. This is so even if, intuitively, the agent keeps her views on causal relations constant — not explicitly changing her mind, that is, about *what causes what.*[15]

This is essential to the role that act counterfactuals play in the Gibbard-Harper formulation of CDT. Recall that according to the Counterfactual Independence Criterion, $Cr^t(a \ \square\!\!\rightarrow \ s)$ and $Cr^t(\neg a \ \square\!\!\rightarrow \ s)$ are set equal to the prior $Cr^t(s)$ *in virtue of the fact* that the agent takes a to have no causal influence over whether s obtains. Because the act-counterfactuals at a time t are in this way set equal to $Cr(s)$ at time t, they are vulnerable to fluctuations via update in $Cr(s)$. In NC Problems $Cr(s)$ is *itself* probabilistically tied to $Cr(a)$, through Conditional Rigidity and the fact that $Cr^t(s \mid a) \gg Cr^t(s)$. The upshot is that rigidity *for counterfactuals* fails over learning experiences on acts:

Fact 2. (Counterfactual non-Rigidity). It is not in general the case that, as an ideal agent undergoes a learning experience on E between any t and $t^+ \geq t$,

$$Cr^t(E \ \square\!\!\rightarrow \phi) = Cr^{t^+}(E \ \square\!\!\rightarrow \phi)$$

---

while on the other hand

$$\begin{aligned}\mathcal{U}_{\text{Egan}}(S) &= \sum_k Cr(S \ \square\!\!\rightarrow \ s_k \ \mid \ S)Val(s_k \wedge S) \\ &= Cr(S \ \square\!\!\rightarrow \ H \ \mid \ S)Val(H \wedge S) \\ &+ Cr(S \ \square\!\!\rightarrow \ \neg H \ \mid \ S)Val(\neg H \wedge S) \\ &= (\textbf{low} \times Val(H)) + (\textbf{high} \times Val(\neg H)).\end{aligned}$$

[15] For example, take their case of King Solomon and Bathsheba (op cit., pg 135). Solomon believes doing unjust things, like sending for Bathsheba (B), would indicate (though not cause) an underlying state, lack of charisma, that foretells revolt (R). They write that "[s]ince [Solomon] knows that B's holding would in no way tend to bring about R's holding, he always ascribes the same probability to $B \ \square\!\!\rightarrow \ R$ as to R" (136). However, $Pr(B \ \square\!\!\rightarrow \ R \mid B) > Pr(B \ \square\!\!\rightarrow \ R)$ (pg. 136): that sending for Bathsheba (B) would result in a revolt (R) is more likely if you *do* send for her.

This contrast distills the real motivation behind Egan's point against classical CDT. The norm of Conditionalization seems to entail:

> (C*) In decision problems, one should anticipate rationally updating act counterfactuals $\ulcorner a_i \boxrightarrow s_k \urcorner$ by conditionalization on one's chosen act.

If (C*) is correct, there would seem to be a strong argument in favor of "Eganized" CDT over classical CDT. The force of Egan's argument comes from the thought that, in situations where an agent expects to get more information as time passes, she should regard her evidence-conditioned credences as better-informed than her current ones. Egan, in effect, asks: should this not be the case for our future credences *in act-counterfactual propositions*, as well as everything else? Assuming an agent in a decision problem is generally self-aware, she can anticipate what her future credence in $(a \boxrightarrow s_k)$ should be, given that she undertakes a.[16] By Conditionalization, this more informed credence is just the *current* Egan credence on s_k in a.[17] Thus reaching for Egan credences, instead of her current act-counterfactual credences, in assessing the utility of acts seems like common sense: an application of Jeffrey's appealing claim that a decision-maker should, in general, "choose for the person [she] expect[s] to be when [she has] chosen" (Jeffrey, 1983, pg.16).

3 The Collapse Lemma

The foregoing is an opinionated — albeit, I think, accurate — account of the state of play vis-a-vis the challenge Egan presents to classical CDT. But there is a serious dialectical puzzle facing that challenge, which becomes clear when we look more closely at the semantics of the counterfactual.

Gibbard & Harper (1978) appeal to just two principles governing the semantics of the '$\boxrightarrow$' connective: Modus Ponens and the Conditional Excluded Middle ("CEM") (Stalnaker, 1968):

$$(A \boxrightarrow S) \vee (A \boxrightarrow \neg S) \qquad \text{(CEM)}$$

It is worth a quick aside to explain why CEM, in particular, is both *controversial from a truth-conditional perspective*—the original context of the Lewis-Stalnaker debate over the

[16] More carefully, by an agent's being "self-aware", we are assuming that *if the agent performs a at t^+, she becomes certain of it: $Cr^{t^+}(a) = 1$.*

[17] Argument: the agent expects that if she brings about a, she will learn (viz., come to have time-t^+ credence 1 that) a. Hence by Conditionalization, her future credence function should be

$$Cr^{t^+}(\cdot) = Cr^t(\cdot \mid a)$$

Plugging in any act counterfactual of the form $a \boxrightarrow s_k$, we conclude that

$$Cr^{t^+}(a \boxrightarrow s_k) = Cr^t(a \boxrightarrow s_k \mid a)$$

...the right-hand side is just the current Egan credence on s_k in a.

truth-conditions of counterfactuals—and *desirable from a probabilistic perspective*, such as Gibbard & Harper's.

First, the controversy. One basis for resistance to CEM is "indeterminate" pairs like

(1) If Bizet and Verdi had been compatriots, Bizet would have been Italian.
$B \mathrel{\square\!\!\rightarrow} I$

(2) If Bizet and Verdi had been compatriots, Bizet would *not* have been Italian.
$B \mathrel{\square\!\!\rightarrow} \neg I$

In light of there being two "equally good" ways for the antecedent to be true — one in which both composers are Italian and one in which both composers are French — some semantic accounts, including the prominent account of Lewis (1973), classify both (1) and (2) (completely) false.[18] Hence both instantiated disjuncts of CEM are false, and CEM itself is no axiom.

But when we move to subjective probability — scoping $\ulcorner a \mathrel{\square\!\!\rightarrow} s \urcorner$ under a credence function $Cr^t(\cdot)$ for the purposes of calculating $\mathcal{U}(a)$ — CEM contributes a nonnegotiable feature: namely, it secures that $Cr^t(a \mathrel{\square\!\!\rightarrow} \cdot)$ is an additive probability function given that $Cr^t(\cdot)$ is. From this it follows that $Cr(a \mathrel{\square\!\!\rightarrow} \neg s)$ goes up *in proportion* to $Cr(a \mathrel{\square\!\!\rightarrow} s)$'s going down.

$$Cr(a \mathrel{\square\!\!\rightarrow} \neg s) = 1 - Cr(a \mathrel{\square\!\!\rightarrow} s) \qquad (\star)$$

This is incompatible with the CEM-rejecting attitude towards (1)-(2) above: the "both (completely) false" view is one on which the agent's credal views concerning *what would happen if she were to perform* a are *sub-additive*: an unacceptable violation of Probabilism.

We return, then, to accepting CEM — at least, for the sake of cashing out CDT, if not for the the sake of cashing out the semantics of natural language counterfactuals.[19] Together, Modus Ponens and CEM entail a principle which Gibbard & Harper call (Consequence 1), and take as the characterizing axiom of the counterfactual:

> **Consequence 1**: $A \supset [(A \mathrel{\square\!\!\rightarrow} B) \equiv B]$
> (Gibbard & Harper, 1978, 127-128).

We are now in a position to articulate the Collapse Lemma from which this paper takes its title. The lemma states that the Egan credence on s in a collapses into the conditional credence in s given a, with the result that for any decision problem and any option $a, \mathcal{U}_{\text{Egan}}(a) = \mathcal{V}(a)$.

[18] These Bizet-Verdi conditionals are based on a examples from Quine (1950), which are much-discussed in Lewis (1973).

[19] Indeed, some authors use something like Equation $(\star)$ to justify rejecting the "both false" intuition in the Bizet-Verdi cases. For example, Stefánsson argues that "the interaction between our confidence[s]" in the counterfactuals like (1)-(2) justifies a truth-conditional semantics that accepts CEM as an axiom (Stefánsson, 2018, §3). For more work on CEM and probability judgments, see Moss (2013b); Eagle (2010); Williams (2012); Mandelkern (2019), and the influential proposals in Skyrms (1980) and Skyrms (1981).

Proof. by the Ratio Formula,

$$\begin{aligned} Cr_{\text{Egan}}(a \text{ in } s) &= Cr(a \boxdot\!\!\rightarrow s \mid a) \\ &= Cr((a \boxdot\!\!\rightarrow s) \wedge a)/Cr(a) \end{aligned}$$

By Consequence 1, applying the biconditional from right to left:

$$\begin{aligned} Cr((a \boxdot\!\!\rightarrow s) \wedge a)/Cr(a) &= Cr(s \wedge a)/Cr(a) \\ &= Cr(s \mid a). \end{aligned}$$

□

Eganized Causal Decision Theory is thus *equivalent* to Evidential Decision Theory: on the leading model-theoretic implementation the '$\boxdot\!\!\rightarrow$' connective, future-directed CDT collapses into classical EDT, a theory that eschews representing causal relationships altogether.

Amongst the immediate dialectical consequences of this result is that it becomes obscure how Egan himself can get the result that one should pick the dominant act in Newcomb's Puzzle. Conceptually, Egan's argument makes it seem like there are three things: (i) the subjective probability of a state, given an act; (ii) the subjective probability that if the act *were* performed, the state *would* result; and (iii) the subjective probability one would have in that same counterfactual, if one learned (only) that the act was actually performed. But it has transpired that (i) and (iii) cannot be distinguished. A causalist-friendly response to the Collapse Lemma, then, is to leverage it to argue that the appearance of there being three things, rather than two, is simply mistaken. The argument from future credence in counterfactuals was just a disguised version of the same reasoning Causalists rightly learned to reject in Newcomb problems. Moreover, the causal decision theorist can provide a complete model theory compatible with this view of counterfactuals.

But looked at another way, the Collapse Lemma is clearly a bizarre result for CDT, too. For the CDTer must confront, not just Egan's particular counterexamples,[20] but his *argument*, which coherently leverages *both* causal notions and concepts related to learning. Given the lemma, these would appear to be on a collision course. The CDTer cannot rely on the Collapse Lemma to deprive the argument of force, since what the proof may *really* indicate is that imposing Consequence 1 on the semantics of counterfactuals issues in a flawed formulation of CDT. Note that dialectically, Egan himself has no reason to endorse Consequence 1. If it fails, and the reduction does not hold, the argument from Murder Lesion can be weakened *from* an argument in favor of $\mathcal{U}_{\text{Egan}}$ (and hence, given the Collapse, in favor of $\mathcal{V}$) to a mere argument *against* $\mathcal{U}$, on grounds that many two-boxers will be tempted to accept. This position — not the endorsement of

[20] There are at least two with *Murder Lesion*'s structure in the original paper. The other widely discussed one is called "Psychopath Button" (Egan op. cit., pg 97).

Eganized CDT, but merely an argument against the classic version of CDT — is indeed Egan's considered view, though for different reasons than the one advanced here.[21]

In the next section, I look at Gibbard & Harper's own treatment of the Collapse Lemma. Readers interested in the connection between the puzzle I have framed for CDT in this section and Lewis's approach to the semantics of counterfactuals, which rejects CEM (Lewis, 1973, 1981), are referred to Fusco (2017); there, I prove that the simplest CDT-friendly way of jettisoning Consequence 1 in fact does little to alter the basic dialectic sketched above.

4 The G&H discussion of the Collapse

Gibbard and Harper derive the basic version of the Collapse Lemma in passing, en route to another point (op. cit., pg. 130). But they return to it later in the paper, with greater emphasis, in a paragraph I reproduce below (first underlined passage). They pair it, as Egan does, with an implicit endorsement of Conditionalization (second underlined passage):

> When a person decides what to do, he has in effect learned what he will do, and so he has new information. He will adjust his probability ascriptions accordingly. These adjustments may affect the $\mathcal{U}$-utility of the various acts open to him.
>
> Indeed, once a person decides to perform an act a, the $\mathcal{U}$-utility of a will be equal to its $\mathcal{V}$-utility. Or at least this holds if Consequence 1...is a logical truth. For we saw in the proof of Assertion 1 that if Consequence 1 is a logical truth, then for any pair of propositions P and Q, $Prob(P \mathrel{\square\!\!\rightarrow} Q \mid P) = Prob(Q \mid P)$. Now let $\mathcal{U}_a$ be the $\mathcal{U}$-utility of an act a as reckoned by the agent after he has decided for sure to do a, and let $Prob$ give the agent's probability ascriptions before he has decided what to do. Let $Prob_a$ give the agent's probability ascriptions after he has decided for sure to do a. Then for any proposition P, $Prob_a(P) = Prob(P \mid a)$. Thus $\mathcal{U}_a(a)$...= $\mathcal{V}(a)$...the $\mathcal{V}$-utility of an act...is what its $\mathcal{U}$-utility would be if the agent knew he was going to perform it. (Gibbard & Harper, 1978, pg. 156-157)

This passage is from pre-Egan times, however, and the result is not viewed by Gibbard & Harper as unsettling. They continue:

> It does not follow that once a person knows what he will do, $\mathcal{V}$-maximization and $\mathcal{U}$-maximization give the same prescriptions. For although for any act a, $\mathcal{U}_a(a) = \mathcal{V}(a)$, it is not in general true that for alternatives b to a, $\mathcal{U}_a(b) = \mathcal{V}(b)$...the distinction between $\mathcal{U}$-maximization and $\mathcal{V}$-maximization remains. (op. cit., pg. 157)

[21] op. cit., pg. 111.

Note here that in Gibbard & Harpers's notation, $\mathcal{U}_c(d)$ (for arbitrary acts c and d) is $\sum_i Cr(s_i \mid c)Val(s_i \wedge d)$; hence in their notation, $\mathcal{U}_a(a)$ what I have called $\mathcal{U}_{\text{Egan}}(a)$.

While I am a partisan of CDT, I am not sure whether, or how, the underlined observation can defend the view from Egan's conceptual argument. Both the quoted passages above begin with a person who has come to *know*, via making up her own mind, that she will do some $a \in \mathcal{A}$. But how did this person decide — and thereby *come* to know — that she was going to do a? At stake is the identity of a — whether it is the $\mathcal{V}$-maximizing act, or the $\mathcal{U}$-maximizing act, of the decision problem the agent faces. (For simplicity, assume the problem is an NC Problem, so that the act must maximize *one* quantity but not both.)

With this in mind, we can re-phrase the observation that $\mathcal{U}_a(a) = \mathcal{V}(a)$ as an Egan-friendly hindsight check. If a is the $\mathcal{V}$-maximizing act (call this "a^E"), then a CDT-like procedure will verify it in hindsight, since the observation that $\mathcal{U}_{a^E}(a^E) = \mathcal{V}(a^E)$ can be glossed as the observation that a^E maximizes $\mathcal{U}$ on the condition that it is decided on. This feature plausibly generates a foresight condition: the agent is in a position to know *prospectively* that *if* she were to choose a^E, it would pass the verification check — meeting with the approval of her "future epistemic self", as Jeffrey might say. If the agent instead chooses the $\mathcal{U}$-maximizing act (call it "a^C"), her act will, by the same token, *fail* the hindsight check. Returning to Egan's example, the person may find herself quite confident that *if she were to shoot, she would hit*. But if shooting itself provides evidence that this counterfactual is really false, and she knows this as she deliberates about whether to shoot, she can anticipate that her high confidence in the counterfactual will be decimated by the evidential impact of taking the shot, reducing $\mathcal{U}(a^C)$ to the already low $\mathcal{V}(a^C)$ (viz., to $\mathcal{U}_{a^C}(a^C)$.)[22]

Hence the interpretive question, in looking at these Gibbard & Harper passages, is this: how does it help to emphasize, as Gibbard & Harper do in the third underlined passage, that "it is not in general true that for alternatives b to a, $\mathcal{U}_a(b) = \mathcal{V}(b)$"? To have a name for both what is granted and what is not, we can set out:

Fact 3. (The Two Faces of Collapse). Although, in any NC Problem, for any act a:
(i) $\mathcal{U}_{\text{Egan}}(a) = \mathcal{U}_a(a) = \mathcal{V}(a)$,
the Causal Expected Utility of a if the agent conditions on a is equal to the prior Evidential Expected Utility of a;

[22] The hindsight check I frame here has much in common with the spirit of *ratifiaibility* of acts (Jeffrey op. cit., pgs 15-16; see also Egan's discussion, pg. 107 ff.). However, it does different dialectical work from the original use of the notion, as can be seen by considering Newcomb's Problem again. Two-boxing is the only ratifiable act in Newcomb's Problem, even though one-boxing maximizes evidential utility from the point of view of Jeffrey-style EDT, because whether the agent comes to condition on $1B$ or on $2B$ (given Collapse, whether the agent becomes nearly certain that $[(1B \boxdot\!\!\rightarrow P1) \wedge (2B \boxdot\!\!\rightarrow P1)]$ or becomes nearly certain that $[(1B \boxdot\!\!\rightarrow P2) \wedge (2B \boxdot\!\!\rightarrow P2)]$), it maximizes evidential expected utility to do $2B$. That dialectic does not apply just as stated to *Murder Lesion*, because it is not the case that whether Mary conditions on shooting *or* she conditions on not shooting, it maximizes her expected utility to shoot. Rather, *Murder Lesion*, like Gibbard & Harper's "Death in Damascus" case, is an example of nontrivial decision dependence (Hare & Hedden, 2015).

(ii) it is not the case that for $a' \neq a$: $\mathcal{U}_a(a') = \mathcal{V}(a')$.
It is not the case that for other acts a', the Causal Expected Utility of a' if the agent conditionalizes on a is equal to the prior Evidential Expected Utility of a'.

In the context of *Murder Lesion*, for example, (ii) is the point that while

$$\mathcal{U}_{\text{shoot}}(\text{shoot}) = \mathcal{V}(\text{shoot}) \tag{1}$$

and

$$\mathcal{U}_{\neg\text{shoot}}(\neg\text{shoot}) = \mathcal{V}(\neg\text{shoot}) \tag{2}$$

are true, the following inequality also holds:

$$\mathcal{U}_{\text{shoot}}(\neg\text{shoot}) \neq \mathcal{V}(\neg\text{shoot}). \tag{3}$$

But it is not clear why this last inequality is of interest.

I'll consider three alternative takes on the underlying dialectic, ending up with one that I favor. As advertised, the third consideration will avert to the peculiarity of predicaments involving (knowingly) choosing one's evidence.

4.1 Take One: The Immediate Post-Act Perspective is Practically Unimportant

A first, simple thought is that even if there is no argument against "pre-conditioning" on one's own acts, this perspective is ultimately unimportant. This is because it is immediately "trumped" by her learning something obviously much more important: viz., learning which $s_k \in \mathcal{S}$ is actual, and thus completely fixing her total payoff for the decision problem.

Most of the cases we have looked at suggest this. Recall again Mary, the agent in *Murder Lesion*. Egan's story emphasizes that as soon as Mary takes the shot, she will instantly have evidence that her shot is unlikely to hit Alfred. If Mary conditionalizes on her act *as* the bullet flies, she is likely to lose hope in a good outcome. But this is obviously a fleeting moment: she is about to *see* whether the bullet hits Alfred. The same dynamic is at work in the classic version of Newcomb's Problem. The way that it is typically told, the choice of either two-boxing or one-boxing leaves little time for epistemic readjustment: the fact of the matter as to whether the big box contains a million dollars, or not, is instantly revealed when the choice is made.[23]

These intermediate moments — *between* the time when an act is chosen and the time when all is revealed — thus occupy an odd position; they loom large in the dialectic that motivates Eganized CDT over classical CDT, but in the context of the cases we've been asked to consider, they seem too ephemeral to be significant loci of epistemic or practical concern. When Mary acts, she is invested in the fate of what we might call her *posterior* future self — that is, the fate of the person who either lives out her days

[23] But see Seidenfeld *op. cit.*, pgs. 204-205.

under tyranny, liberates the nation from Alfred's grip, or is jailed by his cronies. She is *not* directly invested in the features, epistemic or otherwise, of her *proximal* future self — the one who witnesses some temporary fluctuation in attitude towards the utility of her presently available acts.

Alas, this attempt to deflate Eganized causal decision theory — by *practically*, if not exactly *epistemically*, undermining the immediate post-act perspective — is unsuccessful. For the fleetingness of the post-act, pre-revelation period is just an artifact of particular cases. By stretching this period out, allowing plenty of time for the registration of one's regrets, the critical period can be rendered epistemically significant. In a variation on *Murder Lesion*, for example, we can imagine that Mary's option is to lob a javelin at Alfred from an enormous distance. This choice would allow her plenty of time to reflect on her chances of success before "all is revealed".

Moreover, and more potentially embarrassing, a lengthened critical period can be rendered *practically* important, by creating a context in which the agent can *act* on her regrets. If an agent like Mary will predictably regret her choice after she has acted, an enterprising third party can swoop in and offer her — for a fee, of course — a *further* act which will partially offset the consequences of her previous choice in the state she now believes to be most likely. Mary's behavior overall will reflect the self-defeating character of someone who fails to account for what she expected to learn upon acting.

This operationalization represents the logical next step in the evolution of Egan's counterexamples. I know of four similar vignettes in the literature, two due to Ahmed (Ahmed, 2014, Ch. 7.4.3; Ahmed, 2017), one due to Meacham (Meacham, 2010, pg. 64-65) and one due to myself (Fusco, 2018, §3). I provide a simple, *Murder Lesion*-friendly version here, referring the reader to this literature for a more thorough discussion of the way the dialectic interacts with the literature on sequential choice:[24]

> **Murder Lesion, Snailmail Edition**. Mary is deliberating about whether to try to assassinate Alfred by mailing him a bomb. Mary has good evidence that a certain kind of brain lesion, which she may or may not have, causes murderous tendencies but also causes would-be assassins to have significant dyslexia in the writing down of the addresses of their intended victims. This dyslexia is bad enough that if Mary has it, her package is unlikely to reach Alfred, and likely to be delivered to an innocent stranger living somewhere else instead. Mary is currently fairly confident that she does not have mailing-address dyslexia, and not very confident that she will put a bomb in the mail.

We construct a decision matrix that duplicates counterfactual state-descriptions and the payoff relations in the original *Murder Lesion*:

	Mail $\Box\!\!\rightarrow$ Address Incorrect	Mail $\Box\!\!\rightarrow$ Address Correct
Mail	-1000	$+1000$
$\neg$ Mail	0	0

[24] See esp. Ahmed (2014, pg. 204) and Ahmed (2017, §3).

Suppose that, as a classic CDTer, Mary goes ahead and mails the package. She is now quite sure she has the lesion, and thus quite sure her act will have a terrible outcome. I represent this by omitting the ¬ Mail option and shading in the state Mary now assigns the most probability mass to:

	Mail $\Box\!\!\rightarrow$ Address Incorrect	Mail $\Box\!\!\rightarrow$ Address Correct
Mail	-1000	+1000

An entrepreneur now offers Mary a deal: she can pay \$100 for him to intercept her package. Assuming that interception brings about the same results as never having mailed the package at all, taking the deal maximizes causal expected utility from the perspective Mary now occupies:[25]

	Mail $\Box\!\!\rightarrow$ Address Incorrect	Mail $\Box\!\!\rightarrow$ Address Correct
Mail $\wedge$ ¬Deal	-1000	+1000
Mail $\wedge$ Deal	$0-\$100$	$0-\$100$

Looked at as a whole, though — especially if Mary knew, in advance, that someone would offer her the deal if she mailed the package — this course of action is bizarre. Mary has paid \$100 to secure an outcome she could much more easily have guaranteed for free: a life under monotonous tyranny (value: \$0).

I conclude that emphasizing the unimportance of the immediate post-act perspective is not a fruitful way to respond to the Collapse's challenge to CDT.

4.2 Take Two: The Factivity of a

A second possibility for understanding the inequality in Fact 3 is that Gibbard & Harper are drawing attention the factivity of knowledge. Because one cannot know what is false — the thought goes — a deliberating *rational* agent could not genuinely have access *both* to the epistemic position of someone who *has learned* she will do a^E and the epistemic position of someone who *has learned* that she will do a^C. After all, one of these acts is irrational. Hence one quantity, either $\mathcal{U}_{a^E}(a^E)$ or $\mathcal{U}_{a^C}(a^C)$, is not really a rationally accessible causal expected utility: for at least one act $a \in \{a^E, a^C\}$, $U_{\text{Egan}}(a)$ $(=\mathcal{U}_a(a))$ corresponds to the causal expected utility a rational agent would assign to a if

[25] Calculation: suppose for concreteness that initially, Mary's confidence $Cr^t(M \Box\!\!\rightarrow \neg C) = .1$ and $Cr^t(M \Box\!\!\rightarrow C) = .9$. Then $\mathcal{U}(M) = -1000(.1) + 1000(.9) = 800$. However, M is good evidence for $M \Box\!\!\rightarrow \neg C$: $Cr^t(M \Box\!\!\rightarrow \neg C \mid M)$ = (by Collapse) = $Cr^t(\neg C \mid M) = .75$. It follows (since $Cr^t(\cdot)$ is a probability function) that $Cr^t(C \mid M) = .25$. Hence the second offer ("D") will be calculated at:

$$\mathcal{U}(D) = .75(-100) + .25(-100) = -100$$
$$\mathcal{U}(\neg D) = .75(-1000) + .25(1000) = -500$$

Hence $\mathcal{U}(D) > \mathcal{U}(\neg D)$.

she learned something she cannot possibly learn — namely, that she is going to do it. Returning to Jeffrey's maxim to chose for the person you expect to be when you have chosen, the argument could be put like this: one of the two epistemic perspectives afforded by the decision problem — $Cr^t(\cdot \mid a_E)$ or $Cr^t(\cdot \mid a_C)$ — is simply *not* a perspective a rational agent has *any* chance of occupying, once she has chosen.

While this way of cashing out the passage is coherent, it is unsatisfactory as it stands, for two reasons. First, it is a quite general fact about decision-making that agents use conditional probabilities which reflect views on what happens conditional on propositions they regard themselves as having no chance of learning. Suppose that I am faced with an option a that brings some risk of death, and I am spitefully contemplating whether my acquaintances will be remorseful in the event that I die. If this makes a difference to my utilities, then the expected utility of a will depend in part of $Cr(\text{remorse} \mid \text{die})$. This is true even if I regard it as impossible for me learn that I've died.[26]

Second, it isn't clear in general how this dialectical maneuver generalizes to the waxing and waning of credence. In a credal, rather than a full-belief, context, a defender of Egan CDT can accept the letter of factivity while avoiding much of its spirit. On at least *many* views of the latitude we have in deliberation, each epistemic position — knowing that one is going to do a^C, and knowing that one is going to do a^E — can be *approximated*, even if it cannot be fully accepted, by a rational agent in the throes of deliberation.[27] A rational agent who Jeffrey Conditionalizes on a surging resolve to perform a^C, for example, can come arbitrarily *close* to the epistemic position she will occupy if she does a^C.

Indeed, we could simply re-define Egan causal expected utility by appeal to this kind of approximation. The current definition pegs $\mathcal{U}_{\text{Egan}}$ to the $\mathcal{U}$ an agent will assign to a once she has conditionalized on a, which, on a reasonable construal of Conditionalization, is justified only if she comes to *know* a. But Egan could instead have defined it like this:

$$\begin{aligned}\mathcal{U}_{\text{Egan}}(a) &= \lim_{Cr^t(a)\to 1} \mathcal{U}(a) \\ &= \lim_{Cr^t(a)\to 1} \sum_i Cr^t(a \mathrel{\square\!\!\rightarrow} s_i) Val(a \wedge s_i)\end{aligned}$$

On this definition, $\mathcal{U}_{\text{Egan}}(a)$ takes the limit of the classic causal expected utility of a as the agent Jeffrey Conditionalizes on increasing confidence that she will perform a. In all of the example cases that are widely discussed in the literature,[28] this definition yields

[26] This point can also be made with reference to exercises like Jeffrey's *Death Before Dishonor* (Jeffrey, 1983, pg. 89) and, in the semantics literature, by Richmond Thomason's "cheating spouse" examples, discussed by van Fraassen (1980).

[27] On one family of such views, one can "try on" the epistemic perspective of someone who performs an act a by *supposing* that one will do a (Joyce, 2007, §3; Velleman, 1989). This justifies an agent in provisionally increasing his confidence in a.

[28] Including *Murder Lesion*, Egan's *Psychopath Button* (op. cit., pg 97), Gibbard & Harper's *Death in Damascus*, Richter's asymmetric Death in Damascus variant (Richter, 1984, pg. 396).

the same, classical CDT-unfriendly verdicts as the original version of Eganized CDT.

4.3 Take Three: Eganized Credences are Fake News

I proposed above that one way of reading the Gibbard & Harper response to the Collapse is as an argument that one of $Cr_{a^E}(\cdot)$ and $Cr_{a^C}(\cdot)$ represents an illegitimate epistemic perspective for a rational agent. One take on interpreting "legitimacy" had to do with the what can be learned by agents who are rational, and thus cannot make irrational choices. But that interpretation is unpersuasive.

Another way of interpreting epistemic legitimacy — the interpretation I will argue for in the rest of this paper — has to do with evidential quality. This interpretation grants that both a_E and a_C are, in the relevant sense, accessible to the decisionmaker; however, it emphasizes that one of these acts is misleading in respect of which proposition in $\{s, \neg s\}$ is true. The important "causalist" observation is that while, at the moment of decision, the agent can *choose*, by acting, whether her total future evidence will support s or $\neg s$, she cannot choose which of s or $\neg s$ is actually true.

I've thus far assumed the following about an agent, like Mary, facing an NC problem or *Murder Lesion*-like problem (where Gibbard & Harper's rule recommends a^C and EDT/Eganized CDT recommends a^E):

(i) The agent can, in the relevant sense, perform both available acts. Hence both are potentially learnable for her: she can learn a^C (if she does a^C) and she can learn a^E (if she does a^E).

(ii) She will conditionalize on her chosen act.

(iii) Sequential decision problems which exploit the post-act perspective no matter what she does are possible if she does a^C. (Example: the longform version of *Murder Lesion, Snailmail Edition*.)

However, I have not granted the normative upgrade of (ii):

(iv) The agent *ought*, epistemically, to conditionalize on her chosen act.

Indeed, I suggest that (iv) is not generally true. So long as we assume (ii), then, a CDTer has an argument to the effect that an agent who conditionalizes on a^C ends up in a no better (and potentially worse) epistemic position than she occupied before acting.

Before going on to sketch the argument, it is worth being explicit about how, if successful, it affects the dialectic. We seek an account of why causal expected utilities are better calculated according to classical CDT, rather than Eganized CDT. Egan's argument depends on the idea that in decision situations, we *ought* to defer to our future credences. But *if* the CDTer has reason to think her future degrees of belief in $\{s, \neg s\}$ are no better than her current ones, she can reject this call to deference.

To make the case, it will be useful to adopt a time-slice perspective (Hedden, 2015), which conceives of a single persisting agent as an aggregation of different agents at different times. Following Moss (2012), one can take a further step into that perspective by

conceiving of what are ordinarily glossed as diachronic epistemic norms, like Conditionalization, as norms of communication—more particularly, of *testimony and knowledge-transmission*—between the different time-slices that constitute the agent.

Within this structuring metaphor, here are a few platitudes. In general, given that a messenger is statistically reliable with regard to signals $\{e_1, \ldots, e_n\}$ indicating $\{h_1, \ldots, h_n\}$, I should take the messenger's signaling that e_i as evidence that h_i is true. However, this moral must be applied with caution in cases where *I am* the messenger. When I am contemplating different signals $\{e_1, \ldots, e_n\}$ I could send into the ether, I should *not* generally hold that *my* signaling e_i is evidence for *my future self* that h_i is true. This caveat holds even if I have a past track record of being highly reliable on h-related matters.

I take it that the reason for this is no great mystery. My past track record of being a reliable messenger is underwritten by adherence to alethic norms: I generally tried to send only the signal $e_j \in \{e_1, \ldots, e_n\}$ which I antecedently held to be likely on *my own* evidence. In a diachronic context, this means I generally tried to send e_i only when my prior probability for e_i on my total evidence was high. Given my current choice of signals, if I send a signal in e_j which total evidence does not support, I have knowingly flaunted the mechanism which was responsible for my past reliability; I thereby gain no posterior reason to take my signal as novel evidence that h_j is true.

These platitudes can be fruitfully applied in the context a classic NC problem, in which a^C is a signal that is a statistically reliable indicator of $\neg s$. The agent faces a choice between a^C, which immediately secures her a "sweetener" of Δ, and a^E, which foregoes it. The agent can either send her future self good news, or send her future self bad news—"goodness" and "badness" here being understood with respect to whether her act statistically indicates that s is true or false. The sweetener is available just in case the agent sends herself bad news. In only one case, though, will her act constitute a signal to her future self that is non-*misleading* according to her current total evidence: and that is just in case she performs the act which accords statistically with the state-hypothesis (s or $\neg s$) that her *prior* supports.

	s	$\neg s$
a^E	Θ	0
a^C	$\Theta + \Delta$	Δ

The agent located at t, therefore, faces a tradeoff between prudential and (what we might call) testimonial goods. It is a *testimonial* good to refrain from sending misleading evidence to one's future self. (After all, this evidence may be called upon later in future utility-maximization problems.[29]) But it is a *prudential* good to pick up sweeteners while one can. Without a more detailed description of the tradeoff — one which, for example, connects future epistemic states to future utility-maximization problems at particular,

[29]For a classic example of a tradeoff between aiming to maximize expected utility in one's current decision problem and aiming to better one's epistemic position in view of anticipated *future* decision problems, see the "exploration"-"exploitation" tradeoff in Multi-armed Bandit Problems (Robbins, 1952; Berry & Fristedt, 1985).

specified stakes, or states that the agent directly values reliable testimony for its own sake — there is no clear answer to the question of what all-things-considered rationality requires in such a case.[30] All that can be said is that opting for the sweetener is what achieves the *immediate* goal of maximizing utility. And while this *may* be permissible as far as all-things-considered rationality is concerned, it gives the agent no reason to regard her future credences in act counterfactuals — those influenced by suspect testimony — as better than her current ones. Without that presumption, Egan's call to deference to her future credences is blocked.[31]

5 Conclusion

To sum up: we canvassed the collapse lemma itself, and how it is derived from the ratio formula and from commitments regarding the semantics of the counterfactual which are endorsed by Gibbard & Harper's classic account of CDT. I also provided an opinionated account of the lemma's relationship to Egan's counterexample to CDT, as well as Gibbard & Harper's response to the lemma in their 1976 paper.

In closing, it is fruitful to briefly compare the Collapse to Lewis's "Bombshell"—that is, his triviality results for the indicative conditional. In "Probabilities of Conditionals and Conditional Probabilities" (1976), Lewis proved the bizarre result that if $Cr(B \mid A) = Cr(A \rightarrow B)$, then $Cr(A \rightarrow B) = Cr(B)$. Putting the two equalities together and framing the result diachronically, this means conditioning on A does nothing to one's posterior credence in (arbitrary) B. But this is obviously absurd: if learning A at t does anything at all, it changes posterior credences. The weak link in this road to paradox is apparently the commitment about the semantics of the indicative conditional.

For comparison, The Gibbard-Harper collapse result states that, in an NC problem, conditioning $Cr(\cdot)$ on an act a leads to the distinction between causation and correlation's being obliterated in hindsight. Framed diachronically, this means that when agents look backwards, they are insensitive to the difference between knowing they caused s

[30] A similar moral has been emphasized by careful commentators on diachronic Dutch books and other arguments for Conditionalization:

> The claim is not that dynamic coherence and reflection are sufficient for all-things-considered rationality. The claim is that dynamic incoherence and violations of reflection are indicators of *epistemic irrationality*... it is perfectly rational (in the all-things-considered sense) to prefer a situation in which one is slightly epistemically irrational to a situation in which one is perfectly epistemically rational but has to pay all sorts of nonepistemic costs. (Huttegger, 2013, pg. 423; emphasis in original)

[31] Relevant here is Nissan-Rozen (2017), who argues that, in NC problems, the agent's high (even degree-1) credence $Cr(\neg s \mid a^c)$ is *Gettiered* and hence fails to be probabilistic knowledge in the sense of Moss (2013a). Furthermore, Nissan-Rozen claims that *the agent is in a position to know this* about the relevant high conditional credence (*op. cit.*, pg 4813). Whether Nissan-Rozen's view aligns with the diachronic suggestion floated here depends, however, on the question of how agents ought to update in such cases (for example, in the degree-1 case, whether an agent should use Modus Ponens on a conditional she believes, when she *also* knows that her conviction in that conditional is Gettiered.) For more about updating for Causalists, see also Cantwell (2010).

and knowing they merely sent a signal that indicates s without causing it. I think this is as absurd as Triviality: if the causation-correlation distinction does anything at all, it does something which must be capable of being appreciated in hindsight as well as in foresight. Once again, the weak link seems to be a commitment about the semantics of a type of natural language conditional—this time, the counterfactual conditional, rather than the indicative one. But there are many ways for a causal decision theorist to tackle this puzzle, and I have only gestured at one. Confronting this outstanding issue in a model theory of credence is a frontier for formal developments of CDT.

References

Ahmed, Arif (2013). "Causal Decision Theory: A Counterexample." *Philosophical Review* 122. 289–306.

Ahmed, Arif (2014). *Evidence, decision and causality*. Cambridge University Press.

Ahmed, Arif (2017). "Exploiting Causal Decision Theory." Draft, University of Cambridge.

Berry, D. A. & B. Fristedt (1985). *Bandit problems: Sequential allocation of experiments*. Chapman & Hall.

Cantwell, John (2010). "On an alleged counter-example to causal decision theory." *Synthese* 173. 127–152.

Eagle, Anthony (2010). "'Might' counterfactuals." Retrieved through the author's website.

Egan, Andy (2007). "Some Counterexamples to Causal Decision Theory." *The Philosophical Review* 116(1), 93–114.

van Fraassen, Bas (1980). "Review of Rational Belief Systems by Brian Ellis." *Canadian Journal of Philosophy* 10. 497–511.

Fusco, Melissa (2017). "An Inconvenient Proof: The Gibbard-Harper Collapse Lemma for Counterfactual Decision Theory." In Alexandre Cremers, Thom van Gessen & Floris Roelofsen (eds.), *Proceedings of the 21st amsterdam colloquium*, 265–275.

Fusco, Melissa (2018). "Epistemic Time-Bias in Newcomb's Problem." In Arif Ahmed (ed.), *Newcomb's problem*, Cambridge University Press.

Gibbard, Allan & William Harper (1978). "Counterfactuals and Two Kinds of Expected Utility." In C. A. Hooker, J. J. Leach & E. F. McClennen (eds.), *Foundations and applications of decision theory, vol 1*, Dordrecht: D. Reidel.

Hare, Caspar & Brian Hedden (2015). "Self-Reinforcing and Self-Frustrating Decisions." *Noûs* 50(3), 604–628.

Hedden, Brian (2015). "Time-Slice Rationality." *Mind* 124(494), 449–491.

Huttegger, Simon (2013). "In Defense of Reflection." *Philosophy of Science* 80(3), 413–433.

Jeffrey, Richard (1983). *The logic of decision*. University of Chicago Press.

Joyce, James (1999). *The foundations of causal decision theory*. Cambridge University Press.

Joyce, James (2007). "Are Newcomb Problems Really Decisions?" *Synthese* 156. 537–562.

Kment, Boris (2019). "Decision, Causality, and Pre-Determination." Draft, Princeton University.

Lewis, David (1973). *Counterfactuals*. Oxford: Blackwell.

Lewis, David (1981). "Causal Decision Theory." *Australasian Journal of Philosophy* 59(1), 5–30.

Mandelkern, Matthew (2019). "Talking About Worlds." *Philosophical Perspectives* doi: https://doi.org/10.1111/phpe.12112.

Meacham, Christopher (2010). "Binding and its consequences." *Philosophical Studies* 149. 49–71.

Moss, Sarah (2012). "Updating as Communication." *Philosophy and Phenomenological Research* 85(2), 225–248.

Moss, Sarah (2013a). "Epistemology Formalized." *Philosophical Review* 122(1), 1–43.

Moss, Sarah (2013b). "Subjunctive Credences and Semantic Humility." *Philosophy and Phenomenological Research* 87(2), 251–278.

Moss, Sarah (2018). *Probabilistic knowledge*. Oxford University Press.

Nissan-Rozen, Ittay (2017). "Newcomb meets Gettier." *Synthese* 194. 4479–4814.

Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice." In Nicholas Rescher (ed.), *Essays in honor of carl g hempel*, Springer.

Quine, W. V. O. (1950). *Methods of logic*. New York: Holt, Reinhart and Winston.

Richter, Reed (1984). "Rationality Revisited." *Australasian Journal of Philosophy* 62(4), 392–403.

Robbins, H. E. (1952). "Some Aspects of the Sequential Design of Experiments." *Bulletin of the American Mathematical Society* 527–535.

Savage, Leonard (1972). *The foundations of statistics*. Dover.

Skyrms, Brian (1980). *Causal necessity: a pragmatic investigation of the necessity of laws*. Yale University Press.

Skyrms, Brian (1981). "The Prior Propensity Account of Subjunctive Conditionals." In William Harper, Robert Stalnaker & Glenn Pearce (eds.), *Ifs: Conditionals, belief, decision, chance, and time*, Dordrecht: D. Reidel.

Spohn, Wolfgang (1986). "The Representation of Popper Measures." *Topoi* 5.

Stalnaker, Robert (1968). "A Theory of Conditionals." In *Ifs: Conditionals, belief, decision, chance, and time*, D. Reidel, Dordrecht.

Stefánsson, H. Orri (2018). "Counterfactual Skepticism and Multidimensionl Semantics." *Erkenntnis* 83(5), 875–898.

Velleman, David (1989). "Epistemic Freedom." *Pacific Philosophical Quarterly* 70. 73–97.

Williams, J.R.G. (2012). "Counterfactual Triviality: A Lewis-Impossibility Argument for Counterfactuals." *Philosophy and Phenomenological Research* LXXXV(3), 648–670.