The Gibbard-Harper Collapse Lemma for Counterfactual Decision Theory*

Melissa Fusco[1]

Columbia University, New York City, New York, U.S.A.
mf3095@columbia.edu

Abstract

There is a problem for the debate between causal decision theory, formulated in terms of counterfactuals, and its traditional rival, evidential decision theory: an agent's credences in counterfactuals concerning their own acts collapse into evidential probabilities on those acts once diachronic conditionalization on the act is taken into account. Given assumptions that both classical CDTers [6] and their critics (prominently, [4]) accept, it therefore follows that three things cannot be distinct: (i) the probability of a state, given an act; (ii) the probability that if the act *were* performed, the state *would* result; and (iii) the probability one would have in that same counterfactual, if one learned the act was actually performed.

According to both Evidential and Causal decision theory, a choiceworthy act maximizes expected utility. Evidential decision theory (henceforth 'EDT') recommends that one calculate expected utilities using act-conditionalized probabilities on states. Causal decision theory ('CDT') rejects this strategy, with the rationale that some acts—say, seeking out a therapist—do not cause the states—say, being anxious—with which they are correlated. Allan Gibbard and William Harper's "Counterfactuals and Two Kinds of Expected Utility" [6] provides a classic statement of the causal approach, using counterfactual conditionals to express causal relationships that are uniquely of interest to a CDTer.

Many in the recent literature, however, hold that tides have turned against CDT. One important factor is a 2007 paper by Andy Egan, which presents several counterexamples to the theory.[1] On Egan's view, classical causal decision theorists—in particular, those who appeal to the counterfactual formulation of the theory—adhere to the motto "do whatever has the best expected outcome, holding fixed [one's] initial views about the likely causal structure of the world" (96). However, Egan claims, there are situations where agents should not hold such views fixed. In these cases, agents should anticipate their *future* causal views and leverage those instead, thus taking into account what they will learn by performing the act in question. The force of Egan's point is now widely discussed and widely accepted.[2]

In this paper, I focus on an unappreciated formal result, pointed out by Gibbard & Harper [6] in the third section of their classic paper. There, they prove that if an agent is probabilistically coherent, and the semantics for counterfactuals obeys the principle of the Conditional Excluded Middle [24], then the probability of a counterfactual, given its antecedent, collapses to the probability of the consequent, given the antecedent. It follows that "Future-directed Causal Decision Theory" recommends an act just in case classical *Evidential* Decision Theory does. This complicates the dialectic. By Gibbard & Harper's result, the work of putting causal information into a simple EDT system is exactly *undone* by adding a norm of learning. I present a stronger version of the collapse lemma, which applies to Lewis [19]'s weakening of Gibbard & Harper's axioms. I conclude with some thoughts on how to assess the argument from anticipation in light of these results.

*Warm thanks to Andy Egan, Alex Kocurek, Reuben Stern, two anonymous reviewers for the Amsterdam Colloquium, and the audience at the 2015 Formal Epistemology Workshop for discussion and feedback.

[1] These examples were anticipated by Reed Richter in the 1980s: see Richter [21].

[2] See, for example, Arntzenius [1], Briggs [2] and Greaves [7]. For a prominent dissent, see Joyce [13].

1 EDT vs. CDT (in brief)

Both EDT and CDT begin with the idea that the value of each of a set of available acts (call them the a_i's) is calculated by identifying a set of states which fix one's welfare (a set of states s_k) and then multiplying the utility of each state-act conjunction by one's subjective probability, or *credence*, that that state obtains. This relationship is set out by the expected utility equation in Savage's *Foundations of Statistics*,[3] which applies the generic statistical notion of expectation to the agent's welfare function $Val(\cdot)$. For an arbitrary act a,

$$Val_{\text{Savage}}(a) = \sum_k Cr(s_k)Val(s_k \wedge a) \tag{1}$$

The decision-theoretic maxim then enjoins agents to pick the (or an) act a which maximizes the value of this equation.

A concern about this rule is that states sometimes probabilistically depend on acts. To take a common example, suppose you have a test tomorrow and can either study or party tonight. Partying is more valuable than studying, whether you fail or pass. But you would like to pass, and it is worth more to you to pass than to party. In deliberating, it seems you should take into account that $Cr(\mathsf{pass}|\mathsf{study})$—your subjective probability that you'll pass, *given that you study*—is high, while $Cr(\mathsf{pass}|\mathsf{party})$—your subjective probability that you'll pass, *given that you party*—is low.[4] The conditional probabilities $Cr(\mathsf{pass}|\mathsf{study})$ and $Cr(\mathsf{pass}|\mathsf{party})$ are *act-conditionalized* probabilities. Calculating value expectation with these act-conditioned quantities yields

$$Val_{\text{Jeffrey}}(a) = \sum_k Cr(s_k|a)Val(s_k \wedge a) \tag{2}$$

Using Equation 2, rather than Equation 1, may well recommend in this example that you study instead of party—even though partying dominates studying.[5] Jeffrey's equation is often taken to be the defining equation of EDT.

With Equation 2 in hand, utility maximization can be applied to more cases. But not everyone accepts the equation's verdicts in every case to which it can be applied. Causalists, in particular, believe that Jeffrey's equation overgenerates dependencies.[6] To return to a case mentioned above, it seems that one should not avoid seeking out a therapist on the grounds that x's going a therapist is good evidence that x suffers from anxiety, and thus that *ceteris paribus* $Pr(x \text{ is anxious} \mid x \text{ goes to a therapist}) > Pr(x \text{ is anxious} \mid \neg x \text{ goes to a therapist})$. Even if my credences reflect these probabilities for my own case, what matters is that going to the therapist will not *cause* me to be anxious—and I know this.

In light of such considerations, Gibbard and Harper advance a new utility-maximizing equation, wherein the relevant subjective probability is $Cr(a \boxdot\!\!\rightarrow s_k)$. I will call the object of the agent's credence here the *act-counterfactual* $\ulcorner a \boxdot\!\!\rightarrow s_k \urcorner$, and follow Gibbard & Harper in read-

[3] Savage [23].

[4] For reasons of space I have compressed this introduction. Savage himself, of course, cautioned that the simple Equation 1 was not applicable in dependency cases; see Savage [23] and Joyce [12, Ch. 2] for discussion.

[5] Partying dominates studying in the sense that $Val(\mathsf{party} \wedge \mathsf{pass}) > Val(\mathsf{study} \wedge \mathsf{pass})$—you prefer the former in cases where you pass—and $Val(\mathsf{party} \wedge \mathsf{fail}) > Val(\mathsf{study} \wedge \mathsf{fail})$—you also prefer the former in cases where you fail.

[6] It can undergenerate them as well [10], but for ease of exposition I focus on overgeneration here.

ing it as the subjective probability that *if act a were performed, state s_k would obtain*:

$$\mathcal{U}(a) = \sum_k Cr(a \boxdot\!\!\rightarrow s_k) Val(s_k \wedge a) \tag{3}$$

In a "medical Newcomb" problem like the therapist-anxiety case, although $Cr(\mathsf{axious}|\mathsf{therapist})$ is higher than the baseline $Cr(\mathsf{anxious})$, the agent's credence in the act counterfactual, $Cr(\mathsf{therapist} \boxdot\!\!\rightarrow \mathsf{anxious})$, is *not* higher than the unconditionalized credence $Cr(\mathsf{anxious})$, since counterfactuals track causal influence and going to the therapist (by hypothesis) does not *cause* anxiety. Hence dominance reasoning—the same kind of reasoning that seemed *wrong* in the studying case—may now recommend the pursuit of therapy.

2 Egan's Case

While Egan argues that CDT is the wrong theory of decision, he concedes that it gives the right verdict in Newcomb cases [4, pg. 94]. His counterexample to CDT is—at least at a first glance—very different, and goes as follows:[7]

> **Murder Lesion**. Mary is deliberating about whether to shoot Alfred, a loathsome dictator. 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 is currently fairly confident that she has good aim, and not very confident she will shoot.

The basic acts in the Murder Lesion case are: $\{\mathsf{shoot}, \neg\mathsf{shoot}\}$, and the basic states are: $\{\mathsf{hit}, \neg\mathsf{hit}\}$. However, Mary's knowledge includes information about causal influence: she has conditional and unconditional subjective probabilities on well-formed formulas like $(\mathsf{shoot} \boxdot\!\!\rightarrow \mathsf{hit})$—*if I were to shoot Alfred, I would hit him.* Egan argues that the set of states relevant to Mary's decision problem should be, not

(Partition 1) $\{\mathsf{hit}, \neg\ \mathsf{hit}\}$

but

(Partition 2) $\{\mathsf{shoot} \boxdot\!\!\rightarrow \mathsf{hit}, \mathsf{shoot} \boxdot\!\!\rightarrow \neg\ \mathsf{hit}\}$.

Using Partition 2, he argues that the act-conditionalized probability relevant in *Murder Lesion* is given by a more complex formula (*), in which both subjunctive and evidential probability are invoked:

$$Cr(a \boxdot\!\!\rightarrow s_k|a) \tag{*}$$

In Mary's case, it is the credence she has on *if I were to shoot, I would hit, given that I shoot.*

I will henceforth call (*) "the Egan credence on s_k in a." Calculating the expected utility of an act a with Egan credences in state-act pairs yields a quantity we can call $\mathcal{U}_{\text{Egan}}(a)$:

$$\mathcal{U}_{\text{Egan}}(a) = \sum_k Cr(a \boxdot\!\!\rightarrow s_k|a) Val(s_k \wedge a) \tag{4}$$

[7] I have shortened the description somewhat to save space; see Egan [4, pg. 97]. For the addition that Alfred is a dictator, see Joyce [13].

Egan argues that, intuitively, Mary should not shoot in *Murder Lesion*. As Mary confronts her decision, $Cr(\mathsf{shoot})$, by description of the case, is low. Therefore, $Cr(\mathsf{shoot} \boxdot\!\!\rightarrow \mathsf{hit})$ is high, since she is relatively confident she does not have the brain lesion. But her Egan credence $Cr((\mathsf{shoot} \boxdot\!\!\rightarrow \mathsf{hit}) \mid \mathsf{shoot})$ is low, since conditional on shoot, she is quite confident she has the lesion, which causes bad aim. Applying Equation 4, we get Egan's favored answer, which is that shooting has a low expected utility.

Egan's view marks an interesting contrast with classical EDT and CDT. He follows CDT in holding, as against EDT, that causal concepts have a direct role to play in decisionmaking. His argument *assumes* that Mary has credences that explicitly invoke causal influence—and, indeed, that she should reason with these credences as she deliberates.

Why, then, does classical CDT go wrong? What Egan's case adds to the dialectical mix is the diachronic norm of conditionalization. Where $Cr^t(\cdot)$ is an agent's credence function at t and $Cr^t_E(\cdot)$ is the credence function she is disposed to adopt upon learning the proposition E, Conditionalization says

> (Conditionalization) For any proposition A, if an agent learns exactly E between times t and t^+,
>
> $$Cr^{t^+}(A) \stackrel{!}{=} Cr^t(A|E)$$

(I use the symbol '$\stackrel{!}{=}$' to indicate that the equality here is intended to be read normatively, rather than descriptively.)

While it is a normative and not a descriptive claim, Conditionalization does in fact characterize much of our ordinary belief revision. As a norm of betting, it is supported by a Dutch Book argument [26]. Importantly for our purposes, the norm apparently entails, as a special case, that one should update by conditionalization on one's own acts. If this is correct, there would seem to be a strong argument in favor of Egan CDT over classical CDT. The force of Conditionalization comes from the thought that, in situations where an agent expects to get more information as time passes, she should regard her future credences as better-informed than her current ones. Egan, in effect, asks: should this not be the case for our future credences *in act counterfactuals*, as well as everything else? Assuming an agent in a decision problem is generally self-aware, she can anticipate what her credence in $(a \boxdot\!\!\rightarrow s_k)$ will be, given that she undertakes a.[8] By Conditionalization, this future credence is just her *current Egan credence* on s_k in a. Reaching for Egan credences in assessing the utility of acts is therefore enlightened thinking: a straightforward application of Jeffrey's appealing claim that a decision-maker should "choose for the person [she] expect[s] to be when [she has] chosen" [11, pg.16].

3 An Inconvenient Proof: The Collapse Lemma

I think this argument has considerable intuitive force. But my presentation has gotten a bit ahead of itself by neglecting to provide a model theory for counterfactual credence. On the usual order of things, the worldly truth-conditions of well-formed formulas ϕ are prior to their probabilities: $Cr(\phi)$ is the amount of probabilistic mass concentrated on the individual *worlds* where ϕ is true. Gibbard & Harper endorse this order of priority for act counterfactuals (op. cit., 127), taking two axioms to characterize the proposition expressed by $\ulcorner a \boxdot\!\!\rightarrow s_k \urcorner$. The first

[8] More carefully, by an agent's being "self-aware", we are assuming that if the agent does a at t^+, she knows it: $Cr^{t^+}(a) = 1$. See, for example, Hare & Hedden [9, pg. 616]: "you are not prone to astounding yourself."

is Modus Ponens. The second is the *Conditional Excluded Middle* (CEM), which embeds the Law of the Excluded Middle ($S \vee \neg S$) under arbitrary counterfactual antecedents:

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

From these principles, they derive a single characterizing axiom, which they call Consequence 1:

$$A \supset (S \equiv (A \boxrightarrow S)) \tag{Con 1}$$

With this model-theoretic commitment in hand, we can introduce the Collapse Lemma. It uses Consequence 1 to show that the Egan credence on s in a reduces into the conditional credence in s *given* a. The result is that for any decision problem and any option a, $Val_{\text{Jeffrey}}(a) = \mathcal{U}_{\text{Egan}}(a)$. In other words, Equations 4 and 2 are equivalent.
Proof: by the Ratio Formula,

$$\begin{aligned} Cr_{\text{Egan}}(s \text{ in } a) &= Cr(a \boxrightarrow s|a) \\ &= Cr((a \boxrightarrow s) \wedge a)/Cr(a) \end{aligned}$$

By Consequence 1,

$$\begin{aligned} Cr((a \boxrightarrow s) \wedge a)/Cr(a) &= Cr(s \wedge a)/Cr(a) \\ &= Cr(s|a) \\ &\text{QED.} \end{aligned}$$

This is inconvenient for those wishing to press Egan-style counterexamples against CDT: on the leading model-theoretic implementation of the semantic primitives they use, their theory collapses into classical EDT, and so is not a middle-ground position at all. (In particular, it's not clear how Egan can formalize a theory which recommends the dominant option in Newcomb Problems, but also recommends not shooting in *Murder Lesion*.) But the proof is equally inconvenient for CDT. The CDTer must confront, not just Egan's particular counterexamples, but his general *argument*, which brings causal notions into interaction with a norm of learning. It is unsatisfying to rely on Consequence 1 to deprive the argument of force, since the proof may simply indicate that this formulation of CDT is off the mark.

3.1 Strengthening of the Lemma

Where to go from here? An irresistible vantage point on the plausibility of the semantics of counterfactuals is provided by the general framework in Lewis's *Counterfactuals* [17]. According to this familiar, "similarity"-based approach, the counterfactual $\ulcorner A \boxrightarrow S \urcorner$ is true at a world w just in case S is true at all worlds w' which are both A-worlds and *maximally similar* to w.[9]

As many commentators have pointed out, CEM is often rejected in this framework.[10] Sup-

[9] More precisely: where $v' \leq_v v''$ means that v' is more similar to v than v'' is to v, and where, for any proposition p, $max_{\leq,w}(p) := \{w' : w' \in p \wedge \forall w'' : (w'' \in p) \supset (w' \leq_w w'')\}$, the semantics for the counterfactual is: $u \vDash \ulcorner \phi \boxrightarrow \psi \urcorner$ iff $\forall u'$ such that $u' \in max_{\leq,u}$, if $u' \vDash \phi$, then $u' \vDash \psi$. For simplicity, I explicitly consider only the case of atomic ϕ and ψ.

[10] The dialectic is a bit complex here: although Gibbard & Harper invoke CEM, their characterizing axiom, Consequence 1, is in fact weaker than CEM in the similarity framework—it is equivalent instead to Strong Centering [17, pg. 132]. The Collapse Argument relies only on Strong Centering, but because without CEM, $Pr(A \boxrightarrow \cdot)$ is not additive, Gibbard & Harper's appeal the stronger principle is what is plausibly underwriting their commitment to Equation 3 in the first place. For arguments against the weaker (but still strong) Strong Centering principle, see, *inter alia*, Gundersen [8], Leitgeb [14]. For Gibbard & Harper's own reservations about CEM, see op. cit., pg. 128.

pose I am contemplating flipping the coin in my pocket. I know the coin to be fair, so that the outcomes heads and tails would have equal objective probability, or *chance*, of obtaining. It seems strange to hold that the most similar heads-world and the most similar tails-world are not *equally* similar to this world. But CEM allows no ties: it says that either $\ulcorner \mathsf{flipped} \boxright \mathsf{heads} \urcorner$ is true and $\ulcorner \mathsf{flipped} \boxright \mathsf{tails} \urcorner$ is false, or vice-versa. The principle thus breaks the symmetry between chance-symmetric outcomes in an awkward way.

Lewis [20] rejects Gibbard & Harper's formulation of CDT on these grounds, writing that what he calls "the Chance Objection" is "decisive" against CEM [19, pg. 26]. "Fortunately," he adds, "the needed correction is not far to seek." Lewis's alternative version of CDT is probabilistic, trading determinate propositional consequents for determinate chance consequents, like $\ulcorner Ch(\mathsf{heads}) = .5 \urcorner$. Where Pr is a variable over probability functions and Ch is the objective chance function (see 19, pg. 28), Lewis's new equation is:

$$\mathcal{U}_{\text{Fuzzy}}(a) = \sum_k \sum_{Pr} Cr(a \boxright [Ch = Pr]) Pr(s_k) Val(s_k \wedge a) \tag{5}$$

This seems to solve the problem with the coin. Instead of "act-counterfactuals", we might call counterfactuals of the form $\ulcorner a \boxright [Ch = Pr] \urcorner$ (for some probability function Pr) "act-to-chance counterfactuals".

Can act-to-chance counterfactuals provide a framework in which to state Egan's *sui generis* CDT, as distinct from both the classical causal and the classical evidential approach? I don't think so. The problem is that while CEM fails in this framework in a narrow sense, the collapse lemma does not rely on the fact that the consequent of $\ulcorner A \boxright S \urcorner$ is a proposition. Here's a weaker cousin of CEM explicitly tailored to Equation 5:

$$(A \boxright [Ch = Pr]) \vee (A \boxright [Ch \neq Pr]) \tag{CEM-Pr}$$

Since different candidate chance functions exclude each other, if A brings about one chance function, it will exclude any other. And that is what (CEM-Pr) says. This version of CEM is actually *more* plausible than its propositional cousin.

Following Equation 5, let $\mathfrak{P}_{\text{Lewis}}(\cdot)$ be the agent's appropriate counterfactual-on-a credence in s_k, broken down into the best estimate of chance:

$$\mathfrak{P}_{\text{Lewis}}(s_k \text{ in } a) = \sum_{Pr} Cr(a \boxright [Ch = Pr]) Pr(s_k) \tag{P1}$$

Once again, we construct the corresponding Egan credence:

$$\mathfrak{P}_{\text{Egan}}(s_k \text{ in } a) = \sum_{Pr} Cr(a \boxright [Ch = Pr] | a) Pr(s_k) \tag{P2}$$

By a similar proof to the one above, it can be shown that $\mathfrak{P}_{\text{Egan}}(s \text{ in } a) = \mathfrak{P}(s|a)$.[11]

[11] As before:

$$\begin{aligned}
\mathfrak{P}_{\text{Egan}}(s \text{ in } a) &= \sum_{Pr} Cr(a \boxright [Ch = Pr] | a) Pr(s) \\
&= \sum_{Pr} [Cr((a \boxright [Ch = Pr]) \wedge a)/Cr(a)] Pr(s) \\
&= \sum_{Pr} [Cr([Ch = Pr] \wedge a)/Cr(a)] Pr(s) \\
&= \sum_{Pr} Cr([Ch = Pr] | a) Pr(s)
\end{aligned}$$

QED.

We should also note that while (CEM-Pr) looks reasonable, it isn't immediately clear how to model chance consequents in the Lewis similarity framework. In order for it to make sense to say that, in all the most w-similar worlds where A is true, the chance function is (some particular) Pr, one would need to make sense of chances holding at individual worlds, rather than being irreducible features of how probability is spread across the *space* of worlds. Lewis, in his classic statement of chance, is indeed comfortable with worldbound chances, but it is unusual in the typical Bayesian framework.[12]

3.2 Concluding Thoughts

To conclude, I want to return briefly to the question of how a Causalist working in the style of Gibbard & Harper can reply to Egan's conceptual argument about anticipated causal credences.

Our initial puzzle was that, conceptually, it seemed like there were three things: (i) the probability of a state, given an act; (ii) the probability that if the act *were* performed, the state *would* result; and (iii) the probability one would have in that same counterfactual, if one learned the act was actually performed.[13] But it then turned out that (i) and (iii) couldn't be distinguished on a first-pass semantics for the counterfactual. As we saw, the argument initially looks implausible because of its reliance on CEM. But when CEM is weakened in a plausible way, the collapse actually recurs—apparently, with renewed strength.

Therefore, one response to Egan's counterexample is just to rely on the lemma to suggest that the appearance of there being three things, rather than two, was mistaken. The argument from future credence in counterfactuals was just a dressed-up version of the same appeal Causalists learned to reject in Newcomb problems. Moreover, the causal decision theorist can provide a complete model theory compatible with act-to-chance counterfactual view. It is just the Lewisian "closeness" theory, with chances construed as world-bound.

A very different response—one that I favor, but lack the space to fully defend here—also gains strength from the move to chance counterfactuals. But it takes a global, rather than local, perspective on those chance counterfactuals.

Recall that Conditionalization is a norm for getting from one credal state to another, when something is learned. Egan's conceptual argument relied on the strength of this norm. But it is well-known that some sentences in natural language cannot express worldly truth-conditions consistently with this role. Acceptance of such sentences seems to update one's credal state, but it cannot do so by conditionalization.

For example, suppose you learn that *might* A (where 'might' is read epistemically). What should the effect on your credences be? It is hard to believe that the following is *not* an applicable credal norm:

> (Norm-'might') For any proposition A, if the agent learns exactly $\ulcorner$might $A\urcorner$ between times t and t^+,
>
> $$Cr^{t^+}(A) \overset{!}{>} 0.$$

Note here that pooling does not commute with conditionalization, so $\sum_{Pr} Cr([Ch = Pr]|a)Pr(s) \neq \sum_{Pr} Cr([Ch = Pr])Pr(s|a)$ (5, 25, 15).

[12] See e.g. Lewis [16, pg. 269]: "The term 'the chance, at t, of A's holding' is a nonrigid designator...it designates different numbers at different worlds." Once again, coins may be helpful in cashing out such a picture. For example, in any world where I flip Coin 1—regardless of how it lands—there is a certain chance of heads. This chance is different in any world where I flipped the differently weighted Coin 2 instead—even both absolute outcomes, heads and tails, happen in each type of world.

[13] For example, (i) the probability that I am anxious, given that I see a therapist; (ii) the probability that *if I were to see a therapist, it would make me anxious*, and (iii) the probability of (ii), given that I just learned that I *do* see a therapist.

A miniature, two-line "triviality proof" (for example, in the style of 22) can show that (Norm-'might') conflicts with the norm of Conditionalization.[14] Other epistemic expressions ('must', 'probably') and, of course, indicative conditionals (via 18) have the same character.

There is no reason the chance counterfactuals of Equation 5 should not be the same way. On this view, accepting a chance counterfactual puts global constraints on an agent's credences directly—not indirectly, via Conditionalization. Lewis's discussion of probabilistic counterfactual *chances* makes it especially clear how this would proceed: it would go via the Principal Principle [16], the norm that one should conform one's credences to objective chance.

For example, for full belief that one's acts bring about a particular set of chances, we could write the norm as follows:

> (Norm1-$\Box\!\!\rightarrow$) For any state S, if $Cr^t(a \mathbin{\Box\!\!\rightarrow} [Ch = P]) = 1$ and the agent performs act a between times t and t^+,
>
> $$Cr^{t^+}(s) \stackrel{!}{=} P(s)$$

This norm overrides Conditionalization, just as (Norm-'might') does.[15]

This begins to give the CDTer a reply to Egan's conceptual argument. The Causalist's reply should target the underlying appeal to Conditionalization that drives the argument. As other epistemic vocabulary shows, Conditionalization is not a plausible guide to belief revision for *every* well-formed formula. Such formulas are not, however, "lawless"; they simply are governed by more specific normative constraints. In the case of act-to-chance counterfactuals, one could, for example, plausibly substitute (Norm1-$\Box\!\!\rightarrow$) instead.[16]

The claim that counterfactuals generate global constraints on credences, and thus cannot express ordinary, "intersective" propositions, has been made before, on the basis of a somewhat different norm than the one I've advanced here.[17] My concern in this short piece has been

[14] Suppose for reductio that (Norm-'might') is compatible with Conditionalization. Then there is some proposition p (= the one expressed by $\ulcorner$might $A\urcorner$) such that, for any probabilistic credal state $Cr(\cdot)$,

$$Cr(A|p) \stackrel{!}{>} 0$$

Clearly *this* cannot be so, for Cr may rule out A on independent grounds. Begin with an arbitrary probability function $Cr(\cdot)$, and let $Cr'(\cdot)$ be the result of updating Cr with $\neg A$ (viz., so that $Cr'(\cdot) = Cr_{\neg A}(\cdot)$). Now, by Conditionalization,

$$Cr'(A|p) \stackrel{!}{=} Cr(A|\neg A \wedge p) = 0$$

...contradicting (Norm-'might'). See Russell & Hawthorne [22, §3].

[15] For cases of uncertainty—where your confidence is divided between several such Pr—the norm, following Lewis's formulation of $\mathcal{U}_{\text{fuzzy}}$, is:

> (Norm2-$\Box\!\!\rightarrow$) For any state S, if the agent performs act a between times t and t^+,
>
> $$Cr^{t^+}(s) \stackrel{!}{=} \sum_{Pr} Cr^t(a \mathbin{\Box\!\!\rightarrow} [Ch = P])P(s)$$

(Norm2-$\Box\!\!\rightarrow$) entails (Norm1-$\Box\!\!\rightarrow$) as a special case.

[16] Here is a sketch of a mini-Triviality result for (Norm1-$\Box\!\!\rightarrow$). Suppose the agent at t has full belief (credence 1) in $a \mathbin{\Box\!\!\rightarrow} [Ch = Pr']$ for some particular probability function Pr' such that $Pr'(s) = .1$. Moreover, $Cr^t(s) = 0$. The agent performs act a. By the norm, it should be the case that $Cr^{t^+}(s) = .1$. But this cannot be the result of conditionalizing $Cr^t(\cdot)$ on any proposition. The important contrast here is with Lewis [18]'s proof that—given CEM— $Pr(C \mathbin{\Box\!\!\rightarrow} X) = Pr^C(X) = \sum_w Pr(w)Pr(X|w_C)$, where w_C is the unique w-closest C-world (whose existence is guaranteed by CEM).

[17] See especially the discussion in Chapter 6 of Joyce [12], as well as Williams [27] and Briggs [3].

decision theoretic, and thus independent of at least *many* features of the behavior of counterfactuals in natural language. I *do* think it is difficult to defang the argument underlying Egan's counterexamples. But perhaps this is a start.

References

[1] Arntzenius, Frank (2008). "No Regrets, or: Edith Piaf Revamps Decision Theory." *Erkenntnis*, 68(2): pp. 277–297.

[2] Briggs, R.A. (2010). "Decision-Theoretic Paradoxes as Voting Paradoxes." *Philosophical Review*, 119(1): pp. 1–30.

[3] Briggs, R.A. (2017). "Two Interpretations of the Ramsey Test." In Beebee, Hitchcock, and Price (eds.) *Making a Difference: Essays on the Philosophy of Causation*, Oxford University Press.

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

[5] Genest, Christian, and James Zidek (1986). "Combining Probability Distributions: A Critique and an Annotated Bibliography." *Statistical Science*, 1: pp. 114–148.

[6] Gibbard, Allan, and William Harper (1978). "Counterfactuals and Two Kinds of Expected Utility." In Hooker, Leach, and McClennen (eds.) *Foundations and Applications of Decision Theory, Vol 1*, Dordrecht: D. Reidel.

[7] Greaves, Hilary (2013). "Epistemic Decision Theory." *Mind*, 122(488): pp. 915–952.

[8] Gundersen, Lars (2004). "Outline of a New Semantics for Counterfactuals." *Pacific Philosophical Quarterly*, 85(1): pp. 1–20.

[9] Hare, Caspar, and Brian Hedden (2015). "Self-Reinforcing and Self-Frustrating Decisions." *Noûs*, 50(3): pp. 604–628.

[10] Hesslow, Grant (1976). "Discussion: Two notes on the probabilistic approach to causality." *Philosophy of Science*, 43(2): pp. 290–292.

[11] Jeffrey, Richard (1983). *The Logic of Decision*. University of Chicago Press.

[12] Joyce, James (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

[13] Joyce, James (2012). "Regret and Instability in Causal Decision Theory." *Synthese*, 187: pp. 123–145.

[14] Leitgeb, Hannes (2012). "A Probabilistic Semantics for Counterfactuals." *Review of Symbolic Logic*.

[15] Leitgeb, Hannes (2016). "Imaging All The People." *Episteme*, pp. 1–17.

[16] Lewis, David (1971). "A Subjectivist's Guide to Objective Chance." *Studies in Inductive Logic and Probability 2*.

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

[18] Lewis, David (1976). "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review*, 85.

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

[20] Lewis, David (1981). "Why Ain'cha Rich?" *Noûs*, 15: pp. 377–80.

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

[22] Russell, Jeffrey, and John Hawthorne (2016). "General Dynamic Triviality Theorems." *Philosophical Review*, 125(3): pp. 307–339.

[23] Savage, Leonard (1972). *The Foundations of Statistics*. Dover.

[24] Stalnaker, Robert (1981). "A Defense of the Conditional Excluded Middle." In Harper, Stalnaker, and Pearce (eds.) *Ifs: Conditionals, Belief, Decision, Chance, and Time*, D. Reidel.

[25] Steele, Katie (2012). "Testimony as Evidence: More Problems for Linear Pooling." *Journal of Philosophical Logic*, 41: pp. 983–999.

[26] Teller, Paul (1973). "Conditionalization and Observation." *Synthese*, 26: pp. 218–258.

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