Synthese (2022) 200:393
https://doi.org/10.1007/s11229-022-03875-9

ORIGINAL RESEARCH

In defense of causal eliminativism

Alice van't Hoff[1]

Received: 27 October 2021 / Accepted: 26 August 2022 / Published online: 13 September 2022

Abstract
Causal eliminativists maintain that all causal talk is false. The prospects for such a view seem to be stymied by an indispensability argument, charging that any agent must distinguish between effective and ineffective strategies, and that such a distinction must commit that agent to causal notions. However, this argument has been under-explored. The contributions of this paper are twofold: first, I provide a thorough explication of the indispensability argument and the various ways it might be defended. Second, I point to an important limitation in the argument and suggest that it does not give us sufficient reason to reject eliminativism. In support of this last claim, I show that the distinction between effective and ineffective strategies could perfectly well be grounded in a *counterfactual* rather than a *causal* decision theory and argue that there are fully adequate explanations of how we could come to make the requisite counterfactual judgments that need not invoke causal concepts.

Keywords Causation · Counterfactuals · Decision theory · Metaphysics

1 Introduction

Causal eliminativists (henceforth "eliminativists") deny that reality has a causal structure. This is not just a rejection of causal primitivism—the view that the causal relation is an irreducible part of the furniture of reality—but the stronger position that denies that anything answering to our causal talk can be found in the world.[1] A version of this view, asserting that the "law of causality" is false and has no place in science was famously defended by Russell (1913). As I am understanding it here, eliminativism is committed to the ontological claim that a singular causal relation is not actually

[1] Cartwright (1979) endorses something like causal primitivism. Similarly, I understand Michael Tooley's "realism" to constitute a version of the view, Tooley (1990).

✉ Alice van't Hoff
a.vanthoff@princeton.edu

[1] Philosophy Department, Princeton University, 212 1879 Hall, Princeton, NJ 08544, USA

instantiated.[2] I will, however, focus not on the positive case for eliminativism but on the most significant objection faced by its proponents. The contemporary consensus is that eliminativism is false.[3] This is in no small part because, as with other similar projects in philosophy, eliminativists about causation are vulnerable to an indispensability objection that supports *in*eliminativism—the philosophical view that consists in a rejection of eliminativism. I will focus on an argument that suggests that the causal relation is indispensable to the distinctions (which we must draw) between effective and ineffective agential strategies. I will conclude that, initial appearances notwithstanding, this argument does not give rise to a decisive objection against eliminativism.

In Sect. 2, I introduce eliminativism and show how some of the more straightforward pitfalls of the view can be avoided. Section 3 develops the indispensability objection to eliminativism. Sections 4 to 7 suggest a response to that objection. Sections 8 to 9 answer a potential challenge to this response.

2 Dialectic

One version of Russell's central argument goes as follows: it is essential to our concept of causation that only some of an event's antecedents count as causal.[4] Hartry Field puts the point as follows:

> [T]here would be a big deal if we had to conclude that if c_1 and c_2 are both in the past light cone of e then there is no way of regarding one of them as any more a cause of e than the other: then Sam's praying that the fire would go out would be no less a cause than Sara's aiming the water-hose at it, and the notion of causation would lose its whole point. (Field, 2003, p. 439)

In other words, our concept of causation is *discriminating*. It follows that this causal concept could succeed in picking out a relation only if that relation holds uniquely among a select class of relata. Yet there are good reasons to think putatively causal relations cannot be selective in this way. The gist of the argument is that causes must determine their effects, that causes could do so only given a propitious background, that the insurmountable problem of selection means that these background requirements must also be causes, and that, therefore, given the nature of our physical world innumerably many things will count as causes.[5] Our concept of causation therefore

[2] I take it that causal generalizations—"smoking causes cancer"—will also be false if singular causal claims are false, but throughout my focus will only be on singular causation.

[3] E.g. Field (2003). The most popular view in the vicinity is that causation is non-fundamental. Norton (2003) develops a version of this view that has some resemblance to eliminativism, although he explicitly disavows the "eliminativist" label.

[4] For simplicity, I assume throughout that the causal relata are events. This is the mainstream position, but is not unchallenged. It is an assumption in Lewis (1986b) and is defended in Davidson (1980). For alternative views see Bennett (1988), chapters 2–4, 9, Mellor (1995), and Paul (2000).

[5] For an exposition of these ideas see (Russell, 1913, pp. 16–18) and (Field, 2003, pp. 438–440). J. S. Mill offers a famous argument against causal selection—(Mill, 1846, pp. 197–201). Mill's verdict is endorsed in Lewis (1986b), see also (Schaffer, 2016, Sect. 2.3). Lewis (2000) might appear to have the resources to avoid this argument. There he appeals to a notion of *influence*, roughly counterfactual covariation. c then causes e iff c stands in the ancestral of the influence relation to e. We might think that background conditions do not exert a sufficient influence on downstream events to count as causes on this picture. Yet Lewis's

picks out a relation which cannot be instantiated because it is physically or even logically incoherent. Given an Aristotelian view according to which properties exist only when instantiated, there can be no such thing as a causal relation in our world or nearby possible worlds.[6] Of course the eliminativist does not deny that events can be correlated in various complex ways that are manifested in patterns of conditional probabilities. Her contention is just that there is no privileged relation that exists above and beyond these various patterns of interdependence.

In making these claims, the eliminativist relies on a particular interpretation of fundamental science and in particular of physics. Following Russell, she contends that causal concepts are not required to formulate the fundamental laws of physics.[7] This interpretation removes what would be a devastating objection from *physical* indispensability. Nonetheless, other versions of the indispensability objection remain a threat on account of the central role causal notions play in our agential life, both in planning and in the evaluation of action. The canonical version of this idea was presented by Nancy Cartwright. It is worth emphasizing just how important this objection is: it is *the* reference point to which philosophers allude when they set eliminativism aside (Field, 2003, pp. 440–443; Hitchcock, 2007, p. 59; Hitchcock, 2013, p. 139; Price, 2007, pp. 284–288; Woodward, 2007, p. 73). The goal of the next section is to outline Cartwright's argument. Subsequently, I show how an eliminativist might respond.

3 Causal decision theory

According to Cartwright "causal laws cannot be done away with, for they are needed to ground the distinction between effective strategies and ineffective ones" (Cartwright, 1979, p. 420). One of the incontrovertible facts of our world is that there are more and less effective ways of accomplishing certain ends. If it is raining, opening an umbrella is typically a more effective way of remaining dry than not doing so. Some people believe that avoiding vaccinations will be a more effective way of promoting their health over their lifetime than being vaccinated; if we disagree with them, it will be because we think that they are mistaken about the most effective course of action given their goals. Everyone, eliminativists included, should therefore accept:

> **Thesis:** There is a distinction to be drawn between more and less effective strategies.

The goal of the argument from indispensability is to show that this commitment requires that a causal relation is actually instantiated and that some causal claims are true.

Footnote 5 continued
proposal does not avoid the present argument because even if some background factors do not exercise a great influence over the effect, there are innumerably many absences that *do* seem to exert a significant influence.

[6] For a defense of Aristotelianism see Armstrong (1978).

[7] For challenges to Russell's claims see Earman (1976), Suppes (1970). Russell's verdict is endorsed by Field (2003). For further discussion see (Norton, 2003, pp. 3–12), (Hitchcock, 2007, pp. 55–56).

Table 1 Insurance

	Survival chances if not asthmatic (%)	Survival chances if asthmatic (%)
Policyholder	90	50
Non-policyholder	85	40

The critical stage of the ineliminativist's case is a defense of a *causal* decision theory, which can ground the notion of effectiveness required by **Thesis**. The advocates of causal decision theory claim that it is to be favored because it delivers the right verdict in a number of cases that serve as counterexamples to rival analyses. For present purposes, I will assume that these verdicts are correct. Later, I will argue that there are non-causal decision theories available that are consistent with this (concessive) assumption. Consider, then, one of the cases motivating causal decision theory. Suppose that the holders of life insurance policies live longer than non-policyholders without the policy being the cause of their longevity. (Perhaps the kinds of people who purchase life insurance also visit the doctor more frequently, take regular exercise, eat a balanced diet, and have disposable income to invest in healthcare.) Purchasing such policies is not, then, an effective means of promoting a long life, *even though* those who have such policies do in fact live longer. In other words, it is not enough to make a strategy effective that adopting that strategy is merely correlated with the desired outcome (Cartwright, 1979, pp. 429–430). Examples like these give us a preliminary basis for moving towards a causal account of effectiveness.

The problem for eliminativists goes deeper still: it turns out that a strategy might increase the probability of a desired outcome relative to various pieces of a partition while failing to do so overall (or vice versa). This phenomenon is known as "Simpson's paradox". To see how this might arise, suppose again that I am deciding whether to take out a life insurance policy, and imagine that I am either asthmatic or not. I am interested in whether I will make it to 85 years of age. Now suppose that this time I learn that the probability that a policyholder survives to 85 is just 0.58, but the probability that a non-policyholder survives to 85 is 0.76. This makes taking out the policy look foolhardy. But imagine now that, consistent with the aforementioned fact, my chances of survival are as set out in Table 1 below. (This distribution of chances is possible if policyholders are more likely to be asthmatics. In particular, in this case it must be that in the general population of non-policyholders, just one in five people are asthmatic, but four out of every five policyholders have asthma.) It now looks like I should be a policyholder after all. Call this case *Insurance*.[8]

Simpson's paradox shows us that an effective strategy can sometimes even be *negatively* correlated with the desired outcome. While policyholders are less likely to survive until 85 in *Insurance* this is not because the policy has a negative effect on their life prospects, but because it is an indicator of a health condition (asthma) that negatively affects survival. The partition in Table 1 is a more informative basis for decision than the initial probabilities because it holds fixed relevant background factors over which we exert no causal influence. *Insurance* thus suggests that we should not

[8] See Wagner (1982) for real world instances of Simpson's paradox.

accept a decision theory according to which whether an action is rational is a function of how probable desired outcomes are, conditional on the agent performing that action. Theories of this kind are known as "*Evidential Decision Theory*" (EDT). Cases like *Insurance* motivate the replacement of EDT with a *Causal Decision Theory* (CDT).

To see the differences between these views, it will help to express them formally. To do so, we must introduce some notation. Let the S_i partition the space of possible states and take A to be some act whose rationality is in question, then outcomes are given by conjunctions of actions and states $(S_i \wedge A)$. Now let $V(\_)$ be a valuation function assigning a value to action-state conjunctions, and let $Cr(\_)$ be the agent's credence function. Then according to EDT, an action's decision-theoretic value function, $Val_{\mathrm{EDT}}(\_)$, is as follows:

$$Val_{\mathrm{EDT}}(A) = \sum_{i=1}^{n} V(S_i \wedge A) \cdot Cr(S_i \mid A)$$

An action is rational according to EDT to the extent that it realizes a higher decision-theoretic value than the alternatives. By contrast, proponents of CDT hold that to decide what we ought to do we should find the product of the value of an action-state pairing and our credence in the causal hypothesis that the action in question will causally bring about the relevant outcome. Standardly "causal hypothesis" here is cashed out in *counterfactual* terms, with "$\Box\!\!\rightarrow$" standing for a counterfactual connective.[9] Let us call this version of the theory "$\mathrm{CDT_O}$" (for orthodox CDT).[10] $\mathrm{CDT_O}$ evaluates actions as follows:

$$Val_{\mathrm{CDT_O}}(A) = \sum_{i=1}^{n} V(S_i \wedge A) \cdot Cr(A \Box\!\!\rightarrow S_i)$$

As with EDT, a strategy's effectiveness is a matter of its decision-theoretic value relative to other strategies.

While $\mathrm{CDT_O}$, requires that in evaluating what they should do agents must form credences about putatively causal claims, this doesn't in itself require that there be any true causal propositions. One important limitation of the indispensability argument in the present context, then, is that causal notions only feature in causal decision theory within the scope of the credence function. While this observation does draw our attention to a potential weak point in the ineliminativist's case, it need not wholly undermine their argument. Eliminativists maintain that all causal claims are false. So, by their own lights, eliminativists should accord all causal claims credence 0. The point here is not that the rational credence for all agents in causal statements is 0, nor is it that the truth of eliminativism would require all rational agents to assign credence 0 to causal claims (it might very well not, unless eliminativists believe that the evidence supporting eliminativism is wholly *a priori*). Rather, the claim is that eliminativists *themselves* are required to assign credence 0 to causal claims in light of their other

[9] See e.g. (Egan, 2007, p. 95), Gibbard and Harper (1978), (Lewis, 1986a, p. 313), Stalnaker (1981). For discussion of alternative approaches see Lewis (1986a) and Joyce (1999), chapter 5.

[10] $\mathrm{CDT_O}$ was outlined and developed in Stalnaker (1981) and Gibbard and Harper (1978).

doxastic commitments. But then they would no longer be in a position to draw a distinction between effective and ineffective strategies: for all actions and strategies would have the same decision-theoretic value according to their implementation of CDT_O—namely 0. But, the argument continues, it is not enough to satisfy **Thesis** that persons who are *not* eliminativists can continue to draw the distinction between effective and ineffective strategies. Eliminativsts are committed to thinking that they *themselves* can successfully draw such distinctions. Since it seems that they cannot do so, the indispensability argument claims to falsify eliminativism.

4 Eliminativist alternatives to CDT_O

I will suggest that the eliminativist can respond by introducing a decision theory that, like CDT_O, is consistent with the cases developed above to motivate the move away from EDT, but which has no causal commitments (cf. Hall, 2004, pp. 268–269; Hitchcock, 2013, pp. 138–139). Decision theories typically incorporate a dependency condition, offsetting the valuation of an outcome by some measure of an act's tendency to realize this value. To avoid the argument of Sect. 3, the eliminativist must identify some credential state, $Cr(X)$, capturing some kind of action-outcome dependency such that i) for a given act A and state S_i, $Cr(X) \approx Cr(A \mathrel{\Box\!\!\rightarrow} S_i)$ *and*, ii) X comprises no causally committal claims or concepts. i) immediately imposes an important limitation. In virtue of a triviality result proved in Lewis (1986d) (see p. 10 below), $Cr(X)$ cannot be a conditional credence. Instead, I will begin by examining whether the eliminativist can construe X as an acausal counterfactual.

Since this strategy may seem counterintuitive, it is worth proceeding carefully. In the next two paragraphs, I will assume for the sake of argument that a causal relation *is* instantiated. It turns out that, even granted this concession, the dependency condition on which CDT_O relies *cannot* be fully causal if the theory is to get the right results. To see this, consider preemption cases in which causal and counterfactual dependence come apart: if I fatally shoot a man who has just been injected with a deadly poison by an assassin, I cause his death, preempting the process initiated by the assassin. But since the poison would have killed him anyway, it is false that if I hadn't shot the man, he wouldn't have died. So we have a causal relationship, but no counterfactual dependence. A similar phenomenon occurs in cases of overdetermination.[11] Strikingly, where causation and counterfactual dependence come apart, our evaluative judgments seem to track *counterfactual* rather than causal dependencies (Hitchcock, 2013, pp. 138–139). If I am considering whether to strike the ball on the tee with my club, and if doing so would, in the process, deflect away Ada's stroke, which would also have hit the ball, then if I swing, my swinging is the cause of the ball's moving off the tee, but it is nonetheless false that had I not swung, the ball would not have moved. If S is the act of swinging and M is the event of the ball's moving, then $Cr(\neg S \mathrel{\Box\!\!\rightarrow} M)$ should be near 1 (so long as I am rational). But if "$C(\neg S, M)$" represents that $\neg S$ *causes* M then $Cr(C(\neg S, M))$ should be near 0. If all I care about is that the ball

[11] (Paul & Hall, 2013, pp. 74–143) give a comprehensive account of "redundant causation" in counterfactual analyses.

move from the tee and we are using a causal hypothesis as our dependency condition then $Val(S) \approx V(S \wedge M) > Val(\neg S) \approx 0$, whereas if we are using a counterfactual dependency condition $Val(S) \approx V(S \wedge M) \approx Val(\neg S)$. A fully causal decision theory would thus require that I swing, while a counterfactual decision theory implies that both swinging and refraining are rationally permissible.

Fully causal decision theory seems, however, to err here since I can just as well get what I want if I do not swing, given that in this case the process that Ada will initiate no matter what I do would serve to move the ball. Moreover, it isn't clear why I should care about preempting Ada: it is almost fetishistic to insist that *my* action should be the cause of what I care about happening when it would definitely have happened anyway.[12] (There may be further *moral* conditions if, for example, there are agent-relative duties not to act even when your refraining makes no consequential difference. How and how far these can be integrated into the decision-theoretic apparatus is a substantial question in moral philosophy that I will not take up here.) This suggests that it is no coincidence that CDT_O is implemented with a *counterfactual* dependency condition. For, if decision theory is to get the right results in these cases, then what it is rational to do must go by the counterfactual rather than the causal relations that hold (cf. Dorr, 2016, pp. 275–276). In this sense, initial appearances notwithstanding, CDT_O seems to be *misnamed*.

These remarks show how the eliminativist might respond to the indispensability argument: she can defend a *counterfactual* decision theory (CFDT)—$Val_{\text{CFDT}}(A) = \sum_{i=1}^{n} V(S_i \wedge A) \cdot Cr(A \mathrel{\Box\!\!\rightarrow} S_i)$—which, she can argue, provides the best interpretation of the formalism preferred by advocates of CDT_O. Since CFDT reinterprets the proposal of CDT_O it is relatively straightforward to show that the theories agree in their recommendations. The cost of this maneuver is that establishing CFDT's non-causal credentials becomes harder. The next three sections seek to show how this burden could nevertheless be discharged.

5 The entropy theory of counterfactuals

For CDT_O to return the right results in cases like *Insurance*, we must restrict the class of counterfactual judgments involved in decision, setting aside "backtracking counterfactuals"—claims of the form "$A \mathrel{\Box\!\!\rightarrow} C$" that rely on the following kind of reasoning: "if A then B would have had to have already been the case, in which instance C would have followed".[13] To single out the relevant class of judgments, we interpret the counterfactual connective that appears in the formalization of CDT_O in such a way that, according to the official theory, only the right kind of counterfactual judgments are salient to the evaluation of action. The classic Lewis–Stalnaker semantics (Lewis, 1973; Stalnaker, 1968) provides the requisite interpretation, according to which (roughly speaking) "$A \mathrel{\Box\!\!\rightarrow} B$" is true just in case "$B$" is true in all those possi-

[12] If I choose to swing because I want my swing to be the one that strikes the ball, then my action is not a (causal) means to the end, but the end in itself.

[13] Lewis's theory does allow for the occurrence a minor miracle preceding the condition described by the antecedent to feature in our evaluation of counterfactuals (Lewis, 1986c). The important point, though, is that the influence of this miracle should be screened-off by the antecedent.

ble worlds which are such that i) “A” is true at that world, and ii) that world is among the worlds that are most similar to the actual world. When supplemented with an appropriate account of what it is for two worlds to be *similar*—(Lewis, 1986c)—this proposal seems to rule out backtracking interpretations.[14]

Significantly, it turns out that the counterexamples to EDT can be avoided so long as our decision theory eschews a backtracking dependency condition. To see this, consider again *Insurance*. EDT seems to go wrong there because it looks irrational to opt for an action on the basis of its “news value”. We avoid this irrationality by offsetting our valuations of action-state conjunctions by a condition capturing the *forward-looking* connection between action and state, while screening off connections prior to the moment of action. To move beyond EDT we need to find a temporally asymmetric condition capturing this idea. The eliminativist claims that such asymmetries need not be causal; her task is now to show how.

One option is simply to adopt the semantics underpinning CDT$_\text{O}$ lock, stock, and barrel. The ineliminativist, however, can argue against this approach. Counterfactuals can have backtracking readings in a number of contexts; the counterfactual connective that features in CDT$_\text{O}$ is *non*-backtracking because it is intended as a “causal counterfactual” (Lewis, 1986a, p. 326). This requires imposing a number of restrictions on its interpretation. If CFDT is to track the recommendations of CDT$_\text{O}$, its proponents should be able to give a non-arbitrary account of why they use “$\Box\!\!\rightarrow$” to denote only non-backtracking counterfactuals. But, given her rejection of causation, the eliminativist cannot argue that she does so on the grounds that this allows “$\Box\!\!\rightarrow$” to pick out a causal relation. The ineliminativist might then argue that there are no other resources available to eliminativists that could motivate similar restrictions and provide an adequate basis for a non-backtracking reading.

Such claims would be misguided. David Albert and Barry Loewer have developed a physicalistically respectable account of counterfactuals that does not assume any causal notions but supports a non-backtracking reading.[15] Their proposal draws on the way thermodynamic asymmetries are explained in statistical mechanics. One way to characterize the second law of thermodynamics is as a rule to the effect that the entropy of any system that is isolated energy-wise never decreases. The best explanation of this regularity currently relies on what is called the “Past Hypothesis” (PH), which posits a low entropy initial state of the universe at its inception . This hypothesis turns out to be well-supported by contemporary cosmology. We can use PH to formulate (PROB), a probability distribution over the possible initial conditions at some time t that are compatible with PH. If L is a proposition describing the dynamical laws, we can then calculate a statistical mechanical probability function over possible macrostates of the universe—$Pr_{SM}(\_) = \text{PROB}(\_ \mid L \wedge \text{PH})$. If A is a decision taken at some time t, $Pr(\_)$ is a probability function capturing objective chances, and M_t is the macrostate

[14] There are rival proposals, but since the Lewis–Stalnaker view is the most influential I will assume that “$\Box\!\!\rightarrow$” in CDT$_\text{O}$ is to be understood as interpreted by the Lewis–Stalnaker theory.

[15] (Albert, 2000, pp. 125–130) and Loewer (2007); I follow the latter account in my presentation.

of the universe at t, Loewer's proposal is then that:[16]

$$\text{(E)}\ \big(A \mathrel{\Box\!\!\rightarrow} Pr(B) = x\big) \leftrightarrow Pr_{SM}(B \mid A \wedge M_t) = x$$

Call this "the entropy theory of counterfactuals".

While this is a departure from the Lewis–Stalnaker semantics, it turns out to return almost identical results in the cases where they both apply. According to Lewis, a world's similarity to the actual world is a function of similarity in its dynamical laws (and the instances of "miracles"—violations of said laws) and in the size of the region within which the fundamental facts differ from those holding in the actual world. Pr_{SM} is conditional on the dynamical laws holding. States with positive statistical mechanical probability thus realize similarity in this regard to a maximum degree. (This is possible because the macrophysical dynamics turns out to be indeterministic even assuming determinism at the microphysical level.) Moreover, because PH is an *asymmetric* boundary condition, Pr_{SM} is temporally asymmetric in the sense that a decision at t makes a significant difference to the probability of macro events after t, but no difference to the probability of macro events before t. (E) thus avoids backtracking since for any macro event B that did not actually occur prior to the decision A taken at time t, $Pr_{SM}(B \mid A \wedge M_t) = 0$ and so for any such A, we should think that things could not have been macroscopically different before t, if A were (counterfactually) to come to pass.[17]

6 Extending the theory

Unfortunately, (E) is insufficient as a general theory of counterfactuals since it is limited in application to a narrow range of cases. In the present section, I consider which extensions are required if it is to serve the eliminativist's purposes. I will focus first on a restriction that limits the entropy theory to counterfactuals with a particular kind of *antecedent* and second on a limitation that restricts the theory's applicability to counterfactuals with a particular kind of *consequent*. I will argue that the first limitation need not be relaxed, and that while things are different in the second case, there are viable extensions that are consistent with eliminativism.

Loewer's account assumes that the antecedents of the relevant counterfactuals are *decisions*. Clearly though it is possible to assess counterfactuals with different kinds of antecedents. "Had Bradman hit the ball, he would not have been out." True, although the antecedent does not describe a decision. We might therefore wonder whether the entropy-theoretic story can apply to counterfactuals like these. Loewer suggests that it can, although certain restrictions remain. Arguably though this is moot, since such

[16] "Decision" is understood to be a localized microstate of a person's brain, which is correlated with external bodily motion. These assumptions seem permissible given the context. The first is important in avoiding backtracking (Loewer, 2007, pp. 316–317).

[17] An anonymous referee pointed out that this feature of Pr_{SM} seems to rule out non-trivial deliberation about the past, for instance in cases of time-travel. The point is well taken. Since, however, judgments about retro-effectiveness do not feature centrally in our practice of strategy evaluation, eliminativists can arguably set them aside in answering the indispensability objection.

extensions are not required for our purposes. This is important: Dorothy Edgington has argued that it may not be possible to formulate a theory of counterfactuals without causal resources (Edgington, 2004). Her argument is motivated by counterfactuals with antecedents that concern how things were in the *past*. Edgington's analysis is plausible. But her proposal need not disturb the eliminativist. For the counterfactuals that feature in decision theory can feasibly be restricted to cases where the antecedent consists of a future decision. Since Edgington's examples do not fall within this class, they do not undermine the entropy-theoretic account of the counterfactual connective *as it is understood in CFDT*. Cases like those discussed by Edgington might cause problems later for an eliminativist trying to construct a full theory of the world. However, so long as we can develop an eliminativistically acceptable decision theory, the prospects for addressing such problems as they emerge by appeal to what would be an effective means for a hypothetical agent seem good.

A different limitation arises because, in the first instance, Loewer's account handles only counterfactuals whose consequent is a proposition corrresponding to the state of affairs of some event B having probability x. But the counterfactuals that appear in CFDT are of the form: "$A \;\Box\!\!\rightarrow\; S_i$", where S_i is a non-probabilistic state of affairs. The account must therefore be expanded if it is to serve the eliminativist's purposes. One way to generalize it would be to take the ideal credence in "$A \;\Box\!\!\rightarrow\; B$" to be x, where Loewer's account predicts the truth of "$A \;\Box\!\!\rightarrow\; (Pr(B) = x)$" (cf. Kutach, 2002). Given (E) this proposal would make $Cr(A \;\Box\!\!\rightarrow\; B) = Cr(B \mid A \wedge M_t)$. However, as mentioned in Sect. 4, this equality cannot hold generally. Lewis (1986d) proves an important triviality result which suggests that, given certain minimal conditions that are satisfied in the present case, there is no conditional connective "→" such that $Pr(A \rightarrow B) = Pr(B \mid A)$ for all A, B. In these cases $Pr(B \mid A)$ seems to track the *assertibility* (following the terminology of Jackson, 1981) of the conditional "$A \rightarrow B$". Lewis's result thus suggests that probability of truth and assertibility may come apart for conditionals. The Principal Principle implies that credence should follow probability of truth, in which case it should track $Pr(A \rightarrow B)$ rather than $Pr(B \mid A)$ (Lewis, 1986e).

To extend Loewer's account to counterfactuals with non-probabilistic consequents, we must therefore amend (E). Lewis (1986d) shows that while $Pr(A \;\Box\!\!\rightarrow\; B)$ need not equal $Pr(B \mid A)$, there is a way of defining a probability function that is equivalent to $Pr(A \;\Box\!\!\rightarrow\; B)$. To do so, we must introduce the technique of *imaging*. Imagine that we have a credence function defined over a space of possible worlds, such that $Cr_t(w_i)$ reflects our degree of belief at t in the proposition that w_i is actual. Let "$w(X)$" denote a world where X holds. To conditionalize on X, we suppose that at $t_{(i+1)}$ we know that X—that is that our world makes X true. This rules out all worlds where X is false. Conditionalizing on X therefore implies that $Cr_{t_{(i+1)}}(w(\neg X)) = 0$ for all $w(\neg X)$. We then redistribute whatever credence we had previously assigned to worlds where X is false. To do so, we divide $\sum_w Cr_{t_i}(w(\neg X))$ among the worlds where X is true, and attribute this quotient to each such world, producing our credence function for $t_{(i+1)}$. The imaging rule is similar. Again, we suppose that we know that X and eliminate all worlds w such that $w(\neg X)$. This time though instead of redistributing $\sum_w Cr_{t_i}(w(\neg X))$ among all the remaining X worlds, we divide it between only those worlds that are among the most similar to the $\neg X$ worlds (which we are now assigning credence 0).

This suggests a way to refine (E). Let's introduce the notation "$\Box\!\!\rightarrow_{(\mathrm{E})}$" to stand for the counterfactual connective as it is understood by eliminativists and define "$\overline{Pr}^{A}(X \mid Y)$" as the probability function that results from $Pr(X \mid Y)$ after imaging on A. To perform this operation, we need to introduce a similarity metric, but we need not give precise details here. (The account in Lewis Lewis, 1986c is necessarily more finicky and complex, but this is because Lewis's account is unlike ours in relying on the similarity metric to rule out backtracking.) We could then adopt:

$$(\mathrm{E}')\ Cr_t(A \ \Box\!\!\rightarrow_{(\mathrm{E})} B) = \overline{Pr}^{A}_{SM}(B \mid M_t)$$

(Recall that t is the time of decision and M_t is the macrostate at t.) Replacing (E) with (E$'$) removes the second limitation on the entropy theory, allowing it to handle counterfactuals with non-probabilistic consequents.

While (E$'$) bears a superficial resemblance to (E), the differences between them are important: where (E) is a truth-functional proposition, for instance, (E$'$) is an equation of real-valued quantities. We can better understand (E$'$) by introducing the notion of *accuracy*, understood as a graded analog of truth. Accuracy replaces truth as the success condition for variables representing some quantitative state of affairs. (E$'$) specifies when a credence in a counterfactual of the relevant kind is perfectly accurate. Just as Lewis takes the Principal Principle to define the theoretical role of chance as the thing that guides credences (Lewis, 1994, p. 489), so (E$'$) can be understood as picking out "$A \ \Box\!\!\rightarrow_{(\mathrm{E})} B$" as the proposition credence in which is accurate when it agrees with the relevant statistical mechanical probability function. Given the way we have defined $\overline{Pr}^{A}_{SM}(B \mid M_t)$, (E$'$) requires that $Cr(A \ \Box\!\!\rightarrow_{(\mathrm{E})} B)$ is accurate to the extent that it equals the probability of a Lewis–Stalnaker conditional. This is mathematically convenient, as we'll see, but comes at the expense of semantic applicability: since counterfactuals seem to be unlike probability functions in being strictly bivalent—either true or false simpliciter—(E$'$) doesn't give us general semantic conditions for counterfactuals. It turns out, though, that there is broad consensus both that the probabilities of conditionals are interestingly patterned and that these probabilities do not mesh in obvious ways with their truth conditions (e.g. Kaufmann, 2022). Giving a semantics for conditional propositions that makes sense of their probabilities is thus a deep problem in the philosophy of conditionals with which eliminativists need not concern themselves. Their task was to motivate a non-backtracking reading of the counterfactual without resort to causal notions. They accomplished this task by showing that the non-backtracking reading corresponded to the predictions of a physicalistically respectable probability function that tracks entropic asymmetries, and defining "$\Box\!\!\rightarrow_{(\mathrm{E})}$" in terms of this probability function.

The eliminativist can now give their theory of decision:

$$Val_{\mathrm{CFDT}}(A) = \sum_{i=1}^{n} V(S_i \wedge A) \cdot Cr(A \ \Box\!\!\rightarrow_{(\mathrm{E})} S_i)$$

CFDT gives the eliminativist exactly what she wanted: a theory comprising non-backtracking counterfactuals that makes no resort to causal notions. Since it gives the counterfactual connective a non-backtracking reading, CFDT will agree with CDT_O in the cases that motivated causal over evidential decision theory. For the characteristic feature of such cases is that what it is rational to do seems to be responsive to a temporally asymmetric connection between action and outcome, and a non-backtracking counterfactual condition realizes just such a connection. Moreover, there is no arbitrariness in the resulting view, since the proposal has a compelling scientific rationale (Loewer, 2007, p. 320). Counterfactuals *can* be given a backtracking reading, but there is a class of non-backtracking counterfactuals corresponding to an important physical regularity; it is perfectly plausible that this class might be of particular relevance to rational decision. Thus, eliminativists contend that it is CFDT which grounds our judgments about the effectiveness of strategies.

Crucially, agents can apply CFDT without any special knowledge of statistical mechanics; from the perspective of a would-be user of the theory, "$\Box\!\!\rightarrow_{(\text{E})}$" is simply a non-backtracking counterfactual. And while (E$'$) provides a rationale for focusing on non-backtracking readings, it does not give rise to an epistemic standard of evaluation intended to supplement or supplant familiar evidentialist criteria. Rational agents should assign their credences in accordance with their evidence. (E$'$) tells us when such credences would be accurate, but we may always receive misleading evidence that makes it rational to have inaccurate credences. What matters for deliberative purposes, though, is not that our credences are accurate, but that our choices align with the predictions of a decision theory that takes specific credences as inputs. (E$'$) delineates the relevant class of inputs, without adjudicating their rationality.

Consider now a classic Newcomb case (Nozick, 1969). You are offered a choice between two boxes: your options are to pick either just the first box or to pick both boxes. An almost perfectly reliable predictor has predicted what choice you will make. Based on this prediction, the predictor has performed the following action: if they predicted that you will take just a single box, they placed \$1 million in the first box and \$1000 in the second box; if they predicted that you will take two boxes, they placed \$0 in the first box and \$1000 in the second box. Your valuation function is assumed to be directly correlated with your income in dollars. Causal decision theorists argue that the rational thing to do here is to choose both boxes: your decision can make no causal impact on how much money there is in the first box and no matter how much money there is in that box, you do better to take both boxes. CFDT should be able to return this verdict and indeed it does. Consider your choices: "O" or "T", for choosing just one or two boxes respectively. There are two possible states of affairs: in S_1 the first box contains \$1 million, in S_2 it contains \$0. Thus:

$$\begin{aligned}Val_{\text{CFDT}}(O) &= V(S_1 \wedge O) \cdot Cr(O \,\Box\!\!\rightarrow_{(\text{E})} S_1) + V(S_2 \wedge O) \cdot Cr(O \,\Box\!\!\rightarrow_{(\text{E})} S_2)\\ Val_{\text{CFDT}}(T) &= V(S_1 \wedge T) \cdot Cr(T \,\Box\!\!\rightarrow_{(\text{E})} S_1) + V(S_2 \wedge T) \cdot Cr(T \,\Box\!\!\rightarrow_{(\text{E})} S_2)\end{aligned}$$

You do not know whether S_1 or S_2 is the case. But given (E$'$) the counterfactual judgments of relevance to CFDT are non-backtracking. Since the predictor's action precedes your decision, the only counterfactual connections between their action and your decision are backtracking, and so the actual state is counterfactually independent of your choice in the relevant sense of "counterfactual". The rational thing to do is therefore to set your credence in the dependency condition equal to your *unconditional* credence in S_1 or S_2. Suppose you think that there is a 99% likelihood that S_2 obtains and just a 1% likelihood that S_1 is actual (this is a plausible assessment given the description of the case). Then, CFDT predicts:

$$\begin{aligned} Val_{\mathrm{CFDT}}(O) &= 1{,}000{,}000 \cdot 0.01 + 0 \cdot 0.99 = 10{,}000 \\ Val_{\mathrm{CFDT}}(T) &= 1{,}001{,}000 \cdot 0.01 + 1000 \cdot 0.99 = 11{,}000 \end{aligned}$$

Since $Val_{\mathrm{CFDT}}(T) > Val_{\mathrm{CFDT}}(O)$, CFDT recommends taking both boxes.

7 Non-causal counterfactuals

While it seems plausible that causation and counterfactual dependence sometimes come apart, the ineliminativist can argue that a's counterfactually depending on b is sufficient for b to count as a cause of a.[18] In order to settle the issue, we need to say something about what it would take for a relation to count as *causal*. In Sect. 2, I argued for one requirement—that causation must be discriminating in the sense that not all of an event's temporal antecedents can count as causes—and assumed a second—that causes must determine their effects. In what follows, I'll give two arguments against the thesis that counterfactual dependence suffices for causation. The first argument is straightforward but relies on the further assumption that causation must be transitive, while the second argument dispenses with this assumption.

Consider then a family of examples developed by Ned Hall, cases of "double prevention" (Hall, 2004, pp. 241–248). These are cases when one event forestalls a second that would in turn have blocked a third event that actually came to pass. In Hall's version of the case, Suzy is on a bombing mission, but an enemy is sure to shoot her out the sky, preventing her bombing the intended target (B). But Billy, who is piloting a different airplane intervenes (I), attacking the enemy and preventing him from intercepting Suzy. Suzy's bombing counterfactually depends on Billy's intervention, since had Billy not intervened, the enemy would have frustrated her mission. That is: $(\neg I \;\Box\!\!\rightarrow_{(\mathrm{E})} \neg B)$. If counterfactual dependence suffices for causation, then I causes B. In and of itself, this result seems perfectly plausible. Consider though the thesis that causation is transitive in the sense that if x causes y and y causes z, then x ipso facto counts as a cause of z. Ordinary reasoning about causation frequently makes tacit appeal to this thesis: Johnson became President because Lincoln died, Lincoln died because Booth shot him, therefore Booth's shooting Lincoln caused Johnson to become President. Yet if counterfactual dependence suffices for causation and causation is transitive, then absurd conclusions follow. Imagine the enemy is commanded by his superior to intercept Suzy once her incursion is detected (C). Had this command

[18] As observed above, to defend CFDT eliminativists only need to define "$\Box\!\!\rightarrow_{(\mathrm{E})}$" for a limited range of antecedents. In this section, though, I assume that counterfactual defined by the entropy theory can hold more generally, since if there is to be a viable ineliminativist objection here it shouldn't rely on the assumption that the relevant kind of counterfactual dependence is only defined in relation to agential decisions.

not been issued Billy would not have shot down the enemy, so $(\neg C \mathrel{\Box\!\!\rightarrow}_{(\mathrm{E})} \neg I)$. If counterfactual dependence suffices for causation, C causes I and by the reasoning above I causes B. Thus by transitivity, C causes B. But this, Hall shows, is absurd: C initiates a self-canceling threat to B, for if no command is given then the enemy will not begin his sortie and so cannot obstruct Suzy, so $\neg(\neg C \mathrel{\Box\!\!\rightarrow}_{(\mathrm{E})} \neg B)$. C should *not* therefore count as a cause. Thus, if causation is transitive, then by *reductio* counterfactual dependence cannot suffice for causation.

Still the suspicion remains that there are at least some cases where counterfactual dependence *is* enough for causation. Let's relax then the assumption that causation must be transitive and focus on the two other requirements I mentioned above. I suggested first that causes must *discriminate*. This claim is supported by an example of Field's which I discussed in Sect. 2. The intuition is that, conceptually speaking, causation earns its keep as a way of distinguishing among an event's temporal antecedents; some of these antecedents make a special contribution to the event's coming to pass. If our concept of causation were indiscriminate then it would serve no useful function, but it is a necessary condition of concept-possession that the relevant concept should play some cognitive role. Hence causation must be discriminating.

Eliminativists now argue, however, that counterfactual dependence is not discriminating and so cannot be causal. The argument for this claim, developed in Latham (1987), is that (even assuming determinism) it is not possible to give a full specification of the counterfactual conditions of some event e except by specifying *every* parameter in some slice of the back light cone of e. In other words, for any region in e's back light cone there are possible values that would block e's occurrence. e therefore depends counterfactually on every region in any given slice of its back light cone not having any of the e-inhibiting values.[19] On this basis, the eliminativist can claim that counterfactual dependence and temporal precedence are extensionally equivalent—any event in the back light cone of e is equally a temporal antecedent and a potential counterfactual condition of e. It follows that this relation is not discriminating in the requisite sense. Thus, counterfactual dependence does not suffice for causal dependence because it is a weaker relation in virtue of being indiscriminate.

In giving this response, the eliminativist must proceed with caution. For if she is to recover effectiveness judgments she must argue that an agent's available actions can differ in terms of their counterfactual connection with desired outcomes. But this seems, on the face of it, to require that counterfactual dependence *can* be discriminating after all. That is, we might think that discrimination is too stringent a requirement to impose on a causal relation since if eliminativist worlds are *wholly* undiscriminating, then there could be no notion of differential effectiveness in such worlds. Worries of this ilk misunderstand what is meant by "discrimination". To count as undiscriminating, it is not required that a relation R should hold equally between any relata, but only that for any x, y, if x precedes y then $R(x, y)$. The argument of the previous paragraph shows that this is true of counterfactual dependence. That conclusion though is compatible with thinking that eliminativist worlds could have sufficient structure to support differential judgments of effectiveness if the counter-

[19] Assuming that omissions can stand in the relevant relation. This is not only overwhelmingly plausible (see Schaffer, 2000), but also seems undeniable if, as the eliminativist's opponent maintains, counterfactual dependence is causal (cf. Hall, 2004).

factual connections between distinct events can vary in their strength. Significantly, the ineliminativist cannot make an analogous move because causation is a categorical rather than a graded relation: that is, it does not come in degrees but is either "on" or "off" (cf. Kaiserman Kaiserman, 2016 and Sartorio, 2020 here). This helps to make sense of why the indispensability argument does not succeed: differential judgments of effectiveness require only that our world be discriminating in a *graded* sense, while causal structure requires something stronger, namely that the world is discriminating in a categorical sense.

So far, I have said nothing about the second requirement, that causes should determine. The rationale here is straightforward: it is implicit in much of our everyday causal reasoning that if a set of causes $\{c_1, \ldots, c_n\}$ does not suffice to bring about an event e, then there must be some further cause of e—c_{n+1}—missing from our initial set.[20] Although we do not need to rely on the assumption that causes determine to show that counterfactual dependence is not causal, this requirement is nonetheless worth mentioning because it helps us to understand how the resources required for CFDT to have application fall short of those needed for the world to have causal structure. The eliminativist's idea is that discrimination and determination impose conflicting requirements on a relation: a set of discriminating conditions must be exclusive, but the determinants of an event must include all potential defeaters. Not only are judgments of strategy effectiveness not discriminating in a categorical sense, it is also not in general true that effective strategies must necessitate the desired outcome. The structure needed for decision theory to get a grip is thus doubly weaker than that required for the truth of causal claims.

An objector might worry that this picture demands too much of a relation if it is to count as causal. Determination and discrimination seem to me to be relatively *minimal* requirements; still it is worth seeing what the prospects for ineliminativism look like if we relax these demands. One way to do so is to adopt a sort of functionalism. According to this kind of view the success of causal talk in explanation and in decision requires that such talk must be tracking *something*, whatever that thing is will count as causal (cf. Hall, 2004, p. 256). A possible development of this idea is suggested by the "perspectivalist" view, exemplified in Huw Price's work. Price's idea is that causation is a conceptual upshot of our agential perspective. He proposes that c causes e just in case $Pr(e \mid c) > Pr(e)$, where Pr is a probability function calculated *from an agent's point of view*. In making such calculations, an agent supposes that any action available to them is uncaused, originating in their own free decision (Price, 1991, 2007, p. 281; Menzies & Price, 1993, pp. 190–191).

From the eliminativist perspective, however, perspectivalism looks like a kind of *fictionalism* in the sense that perspectivalist judgments are embedded under a condition that we know to be false—namely the supposition that our own actions are independent of whatever diachronic structure subsumes or determines other events. To eliminativists, this is no surprise: notwithstanding their success, our agential and explanatory practices often seem to rest on unrealistic idealizations that belie the inference from the relative success of some practice to the conclusion that the world works

[20] Since the counterfactual conditions of an event need not determine that event, counterfactual theorists of causation may want to resist this characterization—although arguably to do so would be a mistake preventing them capturing our causal concept in full (cf. Kment, 2010, p. 82).

the way the relevant practice represents it as working. This counts strongly against perspectivalist and other quasi-functionalist views. More generally, eliminativists see our causal concept as tainted: they suspect that causal talk is an anthropocentric projection, a legacy of our proto-scientific image of the world. Reasons of conceptual hygiene thus favor the elimination of causal vocabulary.

8 Indispensability revived

Is this sufficient to answer the indispensability argument? It might seem so. But there is an important objection that suggests otherwise: Andy Egan has introduced several cases that give rise to a new version of the argument. Egan presents his cases as objections to CDT_O, however, in light of the discussion above they would seem to be just as accurately described as counterexamples to *CFDT* (Edgington, 2011, p. 80). This suggests a way to resurrect the indispensability argument: so far, I have argued that causal notions are not indispensable to a theory of decision because CFDT is both non-causal and explanatorily adequate. Egan's cases threaten to falsify the second conjunct by showing that a *fully causal* decision theory is required. If my eliminativist response is to succeed, it must be able to head off such cases.

Consider one of Egan's examples: Paul must choose whether or not to press a button that would kill all psychopaths. Paul has a low credence in the proposition that he is a psychopath, and according to his valuation function, it would certainly be better to eliminate psychopathic persons. But Paul believes that no one who was not a psychopath would press the button, and he vastly prefers living in a world of psychopaths to being eliminated himself. Paul, we are inclined to think, should not press the button (Egan, 2007, p. 97). Call this case *Psychopathy*. A plausible specification of Paul's valuation of outcomes is set out in Table 2. Let's stipulate that Paul's counterfactual credences are as follows: $Cr(P \mathrel{\Box\!\!\rightarrow}_{(\text{E})} S_1) < 0.5, Cr(P \mathrel{\Box\!\!\rightarrow}_{(\text{E})} S_2) > 0.5$, $Cr(\neg P \mathrel{\Box\!\!\rightarrow}_{(\text{E})} S_1) < 0.5$, $Cr(\neg P \mathrel{\Box\!\!\rightarrow}_{(\text{E})} S_2) > 0.5$. These stipulations are justified by the description of the case: Paul does not think he is a psychopath, so presumably he should not think he would become one if he presses the button. Hence all the counterfactuals with S_1 in the consequent receive a credence less than 0.5 and all the counterfactuals with S_2 in the consequent receive a credence greater than 0.5. It follows that $10 > Val_{\text{CFDT}}(P) > -20$ but that $-36 > Val_{\text{CFDT}}(\neg P) > -44$. So $Val_{\text{CFDT}}(P) > Val_{\text{CFDT}}(\neg P)$ and so CFDT recommends pressing the button. But that would seem to be the wrong result; there is good reason to think that pressing the button would bring about Paul's death which is, by his own lights, highly undesirable.

Importantly, thinking that Paul would die requires a *backtracking* reading of this counterfactual. For, otherwise, since Paul doesn't initially believe he is a psychopath, and since he shouldn't believe that merely pressing a button can induce psychopathy, he shouldn't believe that pressing the button would kill him. Egan's idea seems to be that our pre-theoretical intuitions support a backtracking reading of this counterfactual and thus that we pretheoretically judge that Paul ought not to press the button. CFDT which places a moratorium on backtracking reasoning must deliver the verdict that Paul would *not* die, were he to press the button, and thus endorses button-pressing. But

Table 2 Psychopathy

	S_1 (Psychopath)	S_2 (Non-psychopath)
P (Press button)	−50	10
$\neg P$ (Do not press button)	−24	−44

this conflicts with our pre-theoretical intuitions, which are, the thought goes, strong enough to support the conclusion that CFDT thereby delivers the wrong verdict.

Edgington suggests that we learn from examples like these that:

> [T]he ban on backtracking is a bad idea. We want all the evidence we can get about what the causal situation will be, *on the assumption that I do* [act] A. In the counterexamples to [EDT] what the backtracking evidence reveals is that there is no causal connection…In the counterexamples to Counterfactual Decision Theory, the backtracking evidence reveals that there is a causal connection. All the examples reinforce that causation is central…(Edgington, 2011, pp. 83–84)

One solution is to move to a *fully* causal decision theory. Instead of credences in counterfactuals we could rely on credences in propositions like: "doing A will cause outcome O" (Edgington, 2011, p. 84). Let's call the resulting view "Edgington Causal Decision Theory" (CDT$_{\text{(E)}}$).[21] Recall that we are using "$C(x, y)$" to indicate that x caused y. CDT$_{\text{(E)}}$ is then the view that:

$$Val_{\text{CDT}_{\text{(E)}}}(A) = \sum_{i=1}^{n} V(S_i \wedge A) \cdot Cr(C(A, S_i) \mid A)$$

CDT$_{\text{(E)}}$ not only avoids problems in *Psychopathy* but can also address the problem posed by *Insurance*. Since having insurance doesn't cause one to live a shorter life, but is merely correlated with a condition that does, conditional on being a policyholder one should have a low credence in the proposition that being a policyholder would reduce your life expectancy. By contrast, since the policy improves the chances of a long life for all, conditional on being a policyholder, you should think it likely that being a policyholder contributes to your longevity. Here, then, we have a way to augment the indispensability argument: Egan's examples show that distinguishing between effective and ineffective strategies requires CDT$_{\text{(E)}}$, rather than CFDT or CDT$_{\text{O}}$. CDT$_{\text{(E)}}$, though, would require eliminativists to assign credences to causal claims that are inconsistent with their view. So eliminativism seems to be falsified.

9 Defending CFDT

One option for the eliminativist would be to reject Egan's account of cases like *Psychopathy*: Paul, she might say, *should* press the button after all. In support of this, the eliminativist could argue that the case is really a twist on a classic Newcomb problem: if it is rational to take both boxes in Newcomb cases, then it should also be rational for Paul to press the button. This is not, though, the best route for the eliminativist. While *Psychopathy* bears some resemblance to a Newcomb case it presents a different

[21] (Edgington, 2011, pp. 84–86). I will depart a little from Edgington's formalism: where she relies on conditional probabilities, I prefer to use conditional credences in causal propositions.

kind of problem. In typical Newcomb cases, the choice preferred by causal decision theorists is supported by dominance reasoning: no matter how the world is, it is better to take both boxes. Things are otherwise in *Psychopathy*: whether P or $\neg P$ is to be preferred depends on whether S_1 or S_2 obtains (cf. Joyce, 2012, pp. 129–130).

Structurally speaking, Newcomb cases provide examples of accidental correlation: some background condition B makes one of the agent's options A and an outcome O covariant, such that the agent's conditional credence $Cr(O \mid A)$ should be relatively high, but B in fact screens off any correlation between O and A. Suppose that O is undesirable, nonetheless intuitively this may not give you reasons against A. Egan's case adds a twist to this structure: B now induces a correlation between A and C, where C is some contextual constraint that *so long as A is performed* in turn makes O more probable. In this structure, the correlation between O and A is not screened off by B. Put differently, the problem in *Psychopathy* isn't that Paul is concerned about pressing the button because that would suggest he might be a psychopath, it is that pressing would suggest that he might be a psychopath *and so* might kill him. Now the badness of O *does* seem to provide a reason against A.

The eliminativist should instead offer a more concessive response to Egan's cases. Decision-theoretic rationality, she can claim, is equivocal: there are conceptions that recommend against pressing the button and other conceptions that recommend pressing. Our evaluative practices do not discriminate between these conceptions. To respond to the indispensability objection, eliminativists should be able to show that they can recover a version of the distinction between effective and ineffective strategies that matches the shape of our practices. But this does not require that they recover *every* version of this distinction. Egan's cases are thus to be corralled to a conception of effectiveness distinct from that which the eliminativist is trying to justify.

In pursuing this approach, the eliminativist can follow the analysis of Egan's cases given in Joyce (2012). Joyce's response starts with the observation that the more your decision favors a certain course of action in this case, the more you should come to believe that it would not be the best thing to do. The choice of P, for example, provides evidence that were the agent to perform P, it would have been better to do something other than P. Since you are an agent facing a decision problem you must think of yourself as responsive to whatever considerations make an option better than its rivals and so your preference for P gives you some reason to revise that very preference. This kind of instability is known as a failure of ratifiability: an action is not ratifiable when choosing that action provides evidence that an alternative option would be the better choice. Neither option in *Psychopathy* is ratifiable (Joyce, 2012, p. 125).[22] It is controversial how and why failures of ratifiability should count against performing a certain action, but one important point seems to be that such failures convey significant information that hadn't previously featured in your evaluation of the options (Joyce, 2012, pp. 138–142; Arntzenius, 2008, pp. 278–280, 290–295).

Joyce (2012) suggests that CDT_O (and so by extension CFDT) permits agents to act only when they are using evaluations that respond to all the relevant information that can be costlessly acquired. Certainly there does seem to be a sense in which someone who chose to act when there was information that was easily available and

[22] (Egan, 2007, pp. 111–113) discusses and rejects several variants of ratificationism.

whose assessment would not impose any costs (so, for example, there is no urgent time pressure) acts irrationally. That's not to deny that there is perhaps another sense in which you do something that's rational when you act on your initial, evidentially unsupplemented, evaluation. But the eliminativist can reasonably deny that this is the version of the effectiveness-ineffectiveness distinction that she is trying to ground. The version to which *she* is committed is a version which carries with it informational requirements.

This last claim may seem somewhat weak until we can see how CFDT *could* satisfy these informational requirements. For if neither P nor $\neg P$ are ratifiable then it's hard to see what an agent could do that would be *better*, and it doesn't seem plausible that there could be decision-theoretic requirements that are in principle impossible to satisfy. Return then to *Psychopathy*. Joyce's suggestion is that we should imagine the agent performing a series of sequential evaluations, updating her credences at each stage. This process can be iterated indefinitely until it reaches a fixed point at which subsequent iterations will not induce any further evaluational changes. A unique equilibrium point of this kind exists in Egan's cases (Joyce, 2012, p. 133). Let t_e denote the time at which the equilibrium point is reached, CFDT can then allow that agents should perform whichever option has the highest valuation according to $Val_{\mathrm{CFDT}^{te}}(\_)$.

Paul must choose whether to press the button. Since any changes to his credence in his performing either P or $\neg P$ will affect his valuation thereof, it must be that his credence at t_e that he will perform P is equal to his conditional credence at t_e that he will perform P given the decision-theoretic evaluation of P at t_e—that is, his credence that he will perform P is unchanged by the information disclosed by the value of $Val_{\mathrm{CFDT}^{te}}(P)$. The same equation must also hold in the case of $\neg P$. It turns out that this is possible at the equilibrium point only if he evaluates both P and $\neg P$ equally. Both are therefore rationally permissible. We now have an explanation of how the informational requirements on CFDT can be satisfied even when neither alternative is ratifiable. The equilibrium valuation incorporates all the information that is revealed by the failures of ratifiability. Even if you act in accordance with the valuations arrived at from the equilibrium point you can anticipate regret once you irrevocably commit to an alternative (since you should then increase your credence in your performing that act to 1). Hence neither option is ratifiable. But once you have reached a decision at equilibrium, the information revealed by failures of ratifiability has already been incorporated into your evaluation. To further revise your decision on these grounds would be a kind of "double counting" (Joyce, 2012, pp. 135–142).

The eliminativist can now finalize her response. CFDT is not subject to Egan-style counterexamples for two reasons: first, because these pertain to a different conception of effectiveness from that which she is concerned to explicate, and second, because (arguably) the ineliminativist was wrong about what a decision theory *should* say in such cases. This also explains why eliminativists are not committed to Edgington's genuinely causal decision theory: CFDT and $\mathrm{CDT}_{(\mathrm{E})}$ deliver different verdicts about Egan's cases. But, there is at least an argument to the effect that it is CFDT that gets things right. CFDT and $\mathrm{CDT}_{(\mathrm{E})}$ are then, at the very worst, on a par. But in that case, $\mathrm{CDT}_{(\mathrm{E})}$ is not indispensable, since there is a perfectly adequate alternative.

Acknowledgements I would like to thank Adam Elga, Mark Johnston, Boris Kment, Harvey Lederman, Gideon Rosen, and Eyal Tal for their help. I also want to thank the Princeton Dissertation Seminar, who read and commented on a draft of this paper, and to extend especial thanks to Haley Brennan for her commentary. Finally, I wish to express my gratitude to three anonymous referees for their suggestions and recommendations. Regrettably, constraints of space prevent me doing full justice to the many objections, insights, and proposals I received from these sources.

Funding No funding was received to assist with the preparation of this manuscript.

Declarations

Competing interest The author has no competing interests to declare that are relevant to the content of this article.

References

Albert, D. Z. (2000). *Time and chance*. Harvard University Press.

Armstrong, D. M. (1978). *Universals and scientific realism* (Vol. 2). Cambridge University Press.

Arntzenius, F. (2008). No regrets, or: Edith Piaf revamps decision theory. *Erkenntnis, 68*, 277–297. https://doi.org/10.1007/s10670-007-9084-8

Bennett, J. (1988). *Events and their names*. Hackett.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs, 13*(4), 419–437. https://doi.org/10.2307/2215337

Davidson, D. (1980). *Causal relations. Essays on actions and events* (pp. 149–162). Clarendon Press.

Dorr, C. (2016). Against counterfactual miracles. *The Philosophical Review, 125*, 241–286. https://doi.org/10.1215/00318108-3453187

Earman, J. (1976). Causation: A matter of life and death. *Journal of Philosophy, 73*(1), 5–25. https://doi.org/10.2307/2025447

Edgington, D. (2004). Counterfactuals and the benefit of hindsight. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world. Cause and chance: Causation in an indeterministic world.* Routledge.

Edgington, D. (2011). Conditionals, causation, and decision. *Analytic Philosophy, 52*(2), 75–87. https://doi.org/10.1111/j.2153-960X.2011.00520.x

Egan, A. (2007). Some counterexamples to causal decision theory. *The Philosophical Review, 116*(1), 93–114. https://doi.org/10.1215/00318108-2006-023

Field, H. (2003). Causation in a physical world. In M. J. Loux & D. W. Zimmerman (Eds.), *The Oxford handbook of metaphysics.* Oxford University Press.

Gibbard, A., & Harper, W. (1978). Counterfactuals and two kinds of expected utility. In A. Hooker, J. J. Leach, & E. F. McClennen (Eds.), *Foundations and applications of decision theory.* D. Reidel.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals.* MIT Press.

Hitchcock, C. (2007). What Russell got right. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited.* Oxford University Press.

Hitchcock, C. (2013). What is the"cause" in causal decision theory? *Erkenntnis, 78*(1), 129–146. https://doi.org/10.1007/s10670-013-9440-9

Jackson, F. (1981). Conditionals and possibilia. *Proceedings of the Aristotelian Society, 81*, 125–137. https://doi.org/10.1093/aristotelian/81.1.125

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

Joyce, J. M. (2012). Regret and instability in causal decision theory. *Synthese, 187*, 123–145. https://doi.org/10.1007/s11229-011-0022-6

Kaiserman, A. (2016). Causal contribution. *Proceedings of the Aristotelian Society, 116*, 387–394. https://doi.org/10.1093/arisoc/aow013

Kaufmann, S. (2022). Bernoulli semantics and ordinal semantics for conditionals. *Journal of Philosophical Logic*. https://doi.org/10.1007/s10992-022-09670-8

Kment, B. (2010). Causation: Determination and difference-making. *Noûs, 44*(1), 80–111. https://doi.org/10.1111/j.1468-0068.2009.00732.x

Kutach, D. N. (2002). The entropy theory of counterfactuals. *Philosophy of Science, 69*(1), 82–104. https://doi.org/10.1086/338942

Latham, N. (1987). Singular causal statements and strict deterministic laws. *Pacific Philosophical Quarterly, 68*, 29–43. https://doi.org/10.1111/j.1468-0114.1987.tb00282.x

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

Lewis, D. (1986a). Causal decision theory. *Philosophical papers* (Vol. 2). Oxford University Press.

Lewis, D. (1986b). Causation. *Philosophical papers* (Vol. 2). Oxford University Press.

Lewis, D. (1986c). Counterfactual dependence and time's arrow. *Philosophical Papers* (Vol. 2). Oxford University Press.

Lewis, D. (1986d). Probabilities of conditionals and conditional probabilities. *Philosophical papers* (Vol. 2). Oxford University Press.

Lewis, D. (1986e). A Subjectivist's guide to objective chance. *Philosophical papers* (Vol. 2). Oxford University Press.

Lewis, D. (1994). Humean supervenience debugged. *Mind, 103*(412), 473–490. https://doi.org/10.1093/mind/103.412.473

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy, 97*(4), 182–197. https://doi.org/10.2307/2678389

Loewer, B. (2007). Counterfactuals and the second law. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited.* Oxford University Press.

Mellor, D. H. (1995). *The facts of causation*. Routledge.

Menzies, P., & Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science, 44*(2), 187–203. https://doi.org/10.1093/bjps/44.2.187

Mill, J. (1846). *A system of logic rationative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Harper & Brothers.

Norton, J. D. (2003). Causation as folk science. *Philosopher's Imprint, 3*(4), 1–22.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In N. Rescher (Ed.), *Essays in honor of Carl G. Hempel.* Reidel.

Paul, L. A. (2000). Aspect causation. *The Journal of Philosophy, 97*(4), 235–256. https://doi.org/10.2307/2678392

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

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science, 42*(2), 157–176. https://doi.org/10.1093/bjps/42.2.157

Price, H. (2007). Causal perspectivalism. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited.* Oxford University Press.

Russell, B. (1913). On the notion of cause. *Proceedings of the Aristotelian Society, 13*, 1–26. https://doi.org/10.1093/aristotelian/13.1.1

Sartorio, C. (2020). More of a cause? *Journal of Applied Philosophy, 37*(3), 346–363. https://doi.org/10.1111/japp.12370

Schaffer, J. (2000). Causation by disconnection. *Philosophy of Science, 67*(2), 285–300.

Schaffer, J. (2016). The metaphysics of causation. In E.N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Vol. 2). Retrieved 26 October, 2021, from https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/.

Stalnaker, R. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory* (pp. 98–112). Basil Blackwell.

Stalnaker, R. (1981). Letter to David Lewis. In W. Harper, R. Stalnaker, & G. Pearce (Eds.), *Ifs.* D. Reidel.

Suppes, P. (1970). *A probabilistic theory of causality*. North-Holland.

Tooley, M. (1990). Causation: Reductionism versus realism. *Philosophy and Phenomenological Research, 50*, 215–236. https://doi.org/10.2307/2108040

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician, 36*(1), 46–48. https://doi.org/10.1080/00031305.1982.10482778

Woodward, J. (2007). Causation with a human face. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality: Russell's republic revisited.* Oxford University Press.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.