Against Counterfactual Miracles

Cian Dorr[1]

Penultimate version. If citing, please consult final version, published in *Philosophical Review* vol. 125, no. 2 (2016), 241–286.

1. The puzzle

There is a puzzle about the truth of counterfactual conditionals in deterministic worlds. It is natural to suppose that if things had gone differently in the recent past—if, say, you had blinked one more time than you actually did while reading the previous sentence—the state of the world in the distant past would still have been exactly as it actually was. It is also natural to suppose that if you had blinked one more time than you actually did, the laws of nature that characterise how the state of the world at earlier times constrains its states at later times would still have been true. But if determinism is true, the combination of the state of the world in the distant past with the laws of nature entails the complete truth about everything that happens, including the truth about how many times you blinked when reading that sentence. So we cannot hold fixed both the laws and the past: something has to give.

We should be precise about the relevant sense of 'determinism'. Let an *intrinsic profile* be an intrinsic property F such that, for any intrinsic property G, either it is metaphysically necessary that everything that has F has G, or it is metaphysically impossible for something to have both F and G. When F is an intrinsic profile that could be instantiated by an interval of time, let F's *history-proposition* be the proposition that some initial segment of the history of the world has F. And let determinism be the claim that every true proposition follows, with metaphysical necessity, from the conjunction of any true history-proposition with all the true laws of nature.[2] Even though there are other things one might mean by 'determinism', determinism in this sense is a live possibility,

one that many physicists and philosophers of physics take quite seriously. So it is not a merely academic exercise to investigate which of our ordinary beliefs are consistent with it.[3]

With this definition in hand, we can state our puzzle more rigorously. Let us stick with the example of blinking, and confine our attention to "normal" people—people without godlike powers, who live long after the beginning of time, in worlds without time travel or other such oddities, and have functional eyelids. Then we can regiment the puzzle as a valid argument from plausible premises to an implausible conclusion:

Counterfactual Consequence Argument (CCA):

Past Necessarily, whenever x is normal at t, there is a true history-proposition p such that p would still have been true if x had blinked at t.

Laws Necessarily, whenever x is normal at t and p is a true law of nature, p would still have been true if x had blinked at t.

Closure Necessarily, whenever x is normal at t and p is metaphysically necessitated by a set of propositions each of which would have been true if x had blinked at t, p would have been true if x had blinked at t.

Triviality Necessarily, whenever x is normal at t and determinism is true, each true proposition would still have been true if x had blinked at t.

The CCA is valid given our definition of determinism. For suppose that x is normal at t and determinism is true. By *Past*, there is a true history-proposition h which would have been true if x had blinked at t. Let h be such a history-proposition, and let S be the set containing h and all true laws of nature. By *Laws*, every member of S would have

been true if x had blinked at t. By determinism, S necessitates every truth. So by *Closure*, every truth would have been true if x had blinked at t.

The *prima facie* appeal of *Past*, *Laws*, and *Closure* is clear. *Triviality*, meanwhile, is deeply implausible. For suppose determinism is true, and that x is some normal creature who fails to blink at t. Then the proposition that x fails to blink at t is true. So according to *Triviality*, this is one of the propositions that would still have been true if x had blinked at t. So if x had blinked at t, x would have failed to blink at t. But of course it is also true that if x had blinked at t, x would have blinked at t. So the upshot of *Triviality* is that the proposition that x blinks at t is *counterfactually impossible*: if it had been true, a contradiction would have been true.[4]

There are four ways out: accept *Triviality*; deny *Closure*; deny *Laws*; deny *Past*. In sections 2–7, I will investigate these options in that order, finally pointing the finger at *Past*. Finally, in section 8 I will consider how reflection on the CCA should affect our evaluation of certain other Consequence Arguments which have been widely discussed in the literature on free will.

Before we begin, we need to take account of an interfering factor which may make it hard to appreciate the difficulty of the puzzle, namely the well-known fact that counterfactual conditionals are context-sensitive. Sentences involving such conditionals generally admit multiple legitimate interpretations, and not all of the premises of the CCA are equally compelling under any interpretation. One particularly relevant example of context-sensitivity is the distinction Lewis (1979) drew between "backtracking" and "standard" contexts. This can be illustrated by an example from Jackson (1977):

(1) a. If I had jumped out of this tenth-floor window, I would have been killed.

b. If I had jumped out of this tenth-floor window, I would have done so only because someone had put a safety net in place.

Assuming I can see that the window overlooks a concrete pavement and am not suicidal, I could legitimately use either (1a) or (1b) to express something well supported by my evidence. But clearly the conjunction of (1a) and (1b) is false on any reasonable interpretation, since it is certainly not the case that if I had jumped out of the window, I would *both* have been killed *and* have jumped because someone had put a safety net in place. So we must be dealing with a case of context sensitivity. Moreover, once we get ourselves into the kind of context evoked by sentences like (1b), *Past* may cease to seem even *prima facie* attractive. We will be tempted to dismiss *Past* on the basis of our reactions to sentences like (2):

(2) If determinism is true and x does not blink at t, then if x had blinked at t, that would have to have been because of a prior history of determining factors differing all the way back.

(2) sounds incontrovertible, and is plausibly true on its most natural interpretation. But it is reasonable to think that (2) might, like (1b), evoke a special kind of interpretation governed by different rules from those that govern the more common kind of interpretation evoked by sentences like (1a).[5] Even if the propositions expressed by *Past* under interpretations of this special kind are obviously false, the propositions expressed by *Past* under more standard interpretations are much more tempting, and have indeed been accepted by many philosophers—mistakenly, in my view.

We could address the issue of context-sensitivity by engaging in metalinguistic ascent, replacing *Laws* and *Closure* with '*Laws*/*Closure* is true in all ordinary contexts',

while replacing *Past* and *Triviality* with '*Past*/*Triviality* is true in many ordinary contexts'. Since valid arguments preserve truth at a context, this still leaves us with a truth-preserving argument with plausible premises and an implausible conclusion. But it will be easier if we allow ourselves to stay at the object level, by stipulating that all counterfactuals are henceforth to be interpreted according to a "standard" interpretation — that is, an interpretation of the kind naturally evoked by sentences like (1a).

2. Accepting *Triviality*

Although embracing *Triviality* is a radical option, it is not a crazy one. Many philosophers, especially in the incompatibilist tradition, have thought that the presupposition that determinism is false is something which all of us — even scientists and philosophers who explicitly embrace determinism — accept at some deep level. If this ingrained assumption is real, one might expect it to be manifested semantically in many ways, perhaps including a pervasive triviality in the truth-values of our counterfactuals relative to worlds where determinism is true.[6]

Nevertheless, *Triviality* should be rejected. Even if we tacitly presuppose indeterminism in our thinking about people, it is not plausible that this somehow infects our thinking about simple deterministic worlds featuring only inanimate objects — billiard balls, dominoes, etc. We have no trouble understanding descriptions of such worlds. Moreover, once we have grasped a sufficiently detailed description, we are very good at using it to assign truth-values to counterfactuals in a non-trivial way ("If seven dominoes were arranged like that, exactly three of them would be such that the bell would ring if they fell"). The mode of reasoning that we use to form judgments about the truth and falsity of counterfactuals on the basis of a description of a scenario in non-counterfactual terms is a genuine skill, and one that works in a broadly similar way whether or

not the scenario we are considering is deterministic. It is not plausible that this mode of reasoning leads us into pervasive error in the deterministic case.

There is one other way of embracing *Triviality* that deserves consideration: one might claim that it is metaphysically impossible for determinism (as I have defined it) to be true, so that *Triviality* is vacuously true. One could try to make this independently plausible by deriving it from the following premises:

Qualitative Laws Only qualitative propositions can be laws of nature.

Haecceitistic Independence Necessarily, there is some non-qualitative truth p and some true history-proposition h such that p is not metaphysically necessitated by h together with the totality of qualitative truths.

Qualitative propositions are those that are not haecceitistic—intuitively, they are not about specific individuals; they say what roles are played but are neutral as regards which objects play which roles. In the language of possible worlds, *Haecceitistic Independence* can be paraphrased as follows: for any possible world w, there is another possible world w' qualitatively indiscernible from w, and haecceitistically indiscernible from w throughout some initial segment of history, but not haecceitistically indiscernible from w *simpliciter*. For example, w and w' might differ merely by a permutation of the qualitative roles played by certain spacetime points in the far future. While this requires a fairly strong form of "haecceitism", it is not all that implausible.[7]

However, this reason for denying the possibility of determinism does not extend to the thesis of 'qualitative determinism': that every true *qualitative* proposition follows, with metaphysical necessity, from the conjunction of any true history-proposition with

all the laws.[8] And for the same reason that *Past*, *Laws*, and *Closure* entail *Triviality*, they also entail the following claim about qualitative determinism:

Qualitative Triviality Necessarily, if qualitative determinism is true, then whenever x is normal at t, each true qualitative proposition would still have been true if x had blinked at t.

Unlike the consequent of *Triviality*, the consequent of *Qualitative Triviality* does not entail that it is counterfactually impossible for any normal person who does not blink at a certain time to do so. Perhaps the population of the world could be counterfactually interlinked in such a way that if any given person x had acted differently, someone else would have instantiated the complete qualitative profile actually instantiated by x. Things might work like this at certain very large and varied possible worlds, such as those where "many worlds" interpretations of quantum mechanics are true (see Dorr n.d.). However, it is not credible that all possible worlds where qualitative determinism is true are like this. If it is possible for qualitative determinism to be true, it is possible for it to be true in a world where there are normal people none of whom ever blinks. And at a world like *that*, it follows from *Qualitative Triviality* that the proposition that no-one ever blinks, being a true qualitative proposition, would still have been true if x had blinked at t (so long as x is normal at t), so that it is counterfactually impossible for x to blink at t. This is surely false.

3. Denying *Closure*

Closure follows from an attractive general principle about counterfactuals:

General Closure If p is metaphysically necessitated by a set of propositions each of which would be true if q were true, then p would be true if q were true.

However, there are reasons to be wary of *General Closure*. Consider the proposition that I am a poached egg. Arguably it is metaphysically impossible, and thus metaphysically necessitates every proposition whatsoever. Since it would be true if I were a poached egg, if we accept *General Closure* we have to say that every proposition would have been true if I were a poached egg. This is odd—for example, the proposition that I am a normal human being does not seem like one that would be true if I were a poached egg.[9]

A different worry about *General Closure* arises in connection with infinite premise-sets. On the influential similarity-based semantics of Lewis (1973), *General Closure* can fail for infinite sets of propositions, although it holds for finite sets.[10] Here is an example using the similarity relation that, according to Lewis (1979), governs the interpretation of counterfactuals in "standard" contexts. Suppose that history is infinite towards the future and finite towards the past, and for any n, let h_n be the proposition that the history of the universe is just as it actually is for the first n seconds. Let p be the proposition that the history of the universe is always just as it actually is. The set of all h_n metaphysically necessitates p. According to Lewis, each proposition h_n is such that if p had been false, it would have been true, since some worlds that match the actual world for the first n seconds are more similar to it than any worlds that do not. But Lewis certainly does not think that if p had been false, p would have been true.[11]

But neither of these worries about *General Closure* is really to the point, since the following weaker principle, which avoids both worries, suffices to justify *Closure*:

Weak General Closure: For any propositions p and q: if p is metaphysically necessitated by a *finite* set of propositions each of which would be true if q were true, *and q is metaphysically possible,* then p would be true if q were true.

First, given how we defined 'normality', it is plausible that if x is normal at t, the proposition that x blinks at t is metaphysically possible; and if not, we could stipulate this as a further condition of normality without thereby making *Triviality* any less implausible. And second, on many accounts of laws of nature the conjunction of any laws will itself be a law, so that in applying *Closure* we need never consider sets with more than two members. Even if this principle is not true in general, resisting *Closure* by appealing to the possibility that there are infinitely many laws of nature is highly unpromising as a strategy for addressing the underlying puzzle: in the kinds of deterministic hypotheses we have reason to take seriously, there are finite sets of laws whose conjunction is a deterministic proposition (that is: a proposition whose conjunction with any history-proposition entails every proposition with which it is consistent).[12]

4. Denying *Past*: can we make do with approximate match?

Given the conclusions of the last two sections, two options remain: we could reject *Past* or reject *Laws*. Both are intuitively unpalatable, but neither is as bad as giving up *Closure* or accepting *Triviality*.

Lewis (1973: 75; 1979) has done most to put the option of rejecting *Laws* on the map. On Lewis's view, if determinism is true, then if you had blinked one more time while reading the first sentence of this paper, the course of history would have been exactly the same until at most a few seconds before you began reading. Then a "small miracle"—a localised exception to some universal generalisation that is actually a law of nature—would have taken place, probably somewhere inside your head, and would

have put the world onto a course where the further exceptionless playing out of the actual laws involved your blinking one more time. While Lewis's ambitious attempt to provide a semantics for counterfactuals from which such judgments can be derived has been highly controversial, the judgments themselves have been widely accepted.[13]

I will argue, contra Lewis, that giving up *Past* is the better option.[14] The present section and section 5 will present and rebut two arguments against the *Past*-denying strategy, the first of which I attribute to Lewis. Sections 6 and 7 will then make a case that the cost of the *Laws*-denying strategy is considerably greater than Lewis and his followers realised.

Lewis's insistence on *Past* is motivated by a genuine insight. When we evaluate an ordinary counterfactual whose antecedent is concerned with a particular interval of time, there is usually some time t—typically only shortly before that interval—such that we tacitly "hold fixed" a very broad range of propositions about history before t, in the sense that we assume that if these propositions are in fact true, they would also have been true in the relevant counterfactual circumstances. When we take ourselves to know one of these propositions, we also treat it as a resource we can draw on in our reasoning about how things would have been different if the antecedent were true. This is how we end up with judgments like the following:

(3) If John had forgotten to have breakfast this morning, that would have been the first time that he did so in months.

(4) If I had been honest during the interview, my colleagues would know that I was fired by my previous employer.

(5) If we convinced a million more people to download this video, we would set a new record.

Call the ordinary pattern of reasoning that generates such judgments *Holding History Fixed*.[15]

A good theory of counterfactuals should, by and large, vindicate the results of *Holding History Fixed*. The alternative is a radical error theory, since so many of our ordinary counterfactual judgments are formed on the basis of reasoning which implicitly appeals to premises about earlier history within the scope of the counterfactual supposition we are developing. However, vindicating the reliability of our ordinary method of forming counterfactual beliefs does not require taking the strict view that if things had gone differently during an interval t, *absolutely all* facts about history before t would have been *exactly* as they actually were. It would be enough to say that history before t would have proceeded *approximately* as it did in the actual world, provided that the relevant sense of "approximately" is one for which we can be confident that most of the historical propositions we actually consider in ordinary applications of *Holding History Fixed* have the same truth value at all worlds whose pre-t history "approximately" matches that of the actual world.[16]

So, what is wrong with this more limited way of vindicating *Holding History Fixed*? Lewis's worry is that there may not *be* any nomically possible worlds which approximately match the actual world up to shortly before t and then diverge in such a way as to make true an antecedent concerning t.[17] Considering the hypothesis that this is how things would have been if Nixon had pressed a certain button at a certain time, he says the following:

> What is worse, there is no guarantee whatever that [a world where the actual laws are true and where Nixon presses the button] can be chosen so that the differences diminish and eventually become negligible in the more and more remote past. Indeed, it is hard to imagine how two deterministic worlds anything

> like ours could possibly remain just a little bit different for very long. There are altogether too many opportunities for little differences to give rise to bigger differences. (Lewis 1979: 45)

Lewis is certainly right that in our universe, little differences tend to blow up quickly into to much bigger differences. But this fact is not a good reason to doubt the existence of nomically possible worlds of the relevant sort. Our best deterministic physical theories have *continuous dynamics*, which means that so long as the past is not infinite, we can always find a nomically possible world which stays arbitrarily close to the actual world throughout any finite initial segment of history, just by choosing an initial state that is close enough to that of the actual world.

This is worth making precise. In each of the theories in question, there is a space M of *dynamical states*, which can be thought of as pairwise inconsistent properties of instants of time. The complete history of the world is entailed by the laws, the facts about which instants there are, their temporal structure (for example, which are before which), and which member of M each one instantiates. M has a certain natural geometrical structure, which includes a topology. There is a *dynamical map* Φ, which is a function (perhaps partially defined) from $\mathbb{R}\times$M (the set of ordered pairs of real numbers and points in M) to M. Φ captures the dynamical content of the laws: the laws entail that whenever the earlier of two instants x units of time apart instantiates p and the later one instantiates p', $p' = \Phi(x, p)$. And Φ is jointly continuous in its two arguments: given a series x_1, x_2, … of real numbers which converges to some x, and a series p_1, p_2, … of members of M which converges to some p, then so long as all members of the series $\Phi(x_1,p_1)$, $\Phi(x_2,p_2)$,… exist, this series converges to $\Phi(x,p)$.

We can regiment Lewis's time-relative notion of similarity between possible worlds using a metric d on M. The distance $d(p,p')$ represents the degree of dissimilarity

between w at t and w' at t' when t instantiates p at w, t' instantiates p' at w', and both w and w' are nomically possible. It is plausible that d is continuous: when p_1, p_2, … converges to p and q_1, q_2, … converges to q, $d(p_1,q_1)$, $d(p_2,q_2)$,… converges to $d(p,q)$. Combined with the continuity of Φ, this has the following consequence: for every $p \in M$ and positive real numbers t and ε, there is an open neighbourhood O of p such that for every $p' \in O$ and $x \in [0,t]$, $d(\Phi(x,p'),\Phi(x,p)) < \varepsilon$.[18] In other words, by choosing an initial state p' close enough to the actual initial state p, we can ensure that the world remains as similar as we like to the actual world for as long as we like.

Of course, the fact that there are nomically possible worlds that stay very similar to actuality until shortly before t but diverge after t does not by itself establish that there are nomically possible worlds of the kind that Lewis was worried about—for example, worlds that stay very close to actuality until shortly before t and at which Nixon goes on to press the button at t. Could getting t into the particular region of the state space it needs to occupy for Nixon to press the button require a trajectory that diverges substantially from actuality long before t, so that many ordinary sentences about history before t get different truth values? In principle the answer could be yes, but it is extremely unlikely. The key to seeing why is the fact emphasised by Lewis, that little differences characteristically blow up quickly into much bigger differences. On any reasonable similarity metric d on M, typically, when p is a state instantiated by some time at the actual world and q is some other state such that $d(p,q)$ is small, and x is (say) one second, $d(\Phi(x,p), \Phi(x,q))$ is some large multiple of $d(p,q)$. Thus, if we take the set of nomically possible worlds that stay within ε of the actual world until t, and look at the set of dynamical states instantiated at these worlds by the time one second after t, we will find this set scattered over a much wider region of M, in such a way that the mean distance between them is some large multiple of ε. The faster small differences blow up into big

ones, the more scope there is for wide variation in what happens shortly after t among the worlds that cleave tightly to actuality until t.

Extreme sensitivity to small differences is one hallmark of *chaotic* systems. These systems have dynamical maps that, when given a small volume of M, exponentially stretches it out in some directions. If the initial region was a sphere (according to some reasonable metric), then its image under Φ will soon be a long, thin, branching strand, squiggling back and forth in an intricate way. Even though the volume of the image is the same as the volume of the initial region (in classical dynamics, M carries a natural volume measure which is preserved by the dynamics), because the image is so stretched out, the mean distance between its points, and the volume of the set of points that are *close* to it, are both much greater than those of our original sphere. The study of chaotic dynamical systems is a thriving topic within classical mechanics. This literature strongly confirms Lewis's claim that such sensitive dependence is widespread. It crops up even in some very simple models with just a few degrees of freedom; and in general, we expect that any complex system with many interacting degrees of freedom will display such sensitivity to a high degree.[19]

The prevalence of chaos suggests in a rather abstract way that we should expect the set of worlds that approximately match actuality until t to be quite varied as regards history after t. To gain a more concrete understanding of the kinds of variations we can expect to find within this set, let us consider a famous example from statistical mechanics. Suppose that at t, the air in the seminar room is at thermodynamical equilibrium. Then, using a kind of reasoning which plays a central role in statistical mechanics, we can assure ourselves that there are some nomically possible worlds which approximately match actuality until t, at which the air molecules subsequently move in such a way that at $t+\Delta t$ they end up congregated in one tiny corner of the room (where Δt is

the short amount of time it would typically take for the molecules, once squeezed into the corner, to explode back out to equilibrium).

To convince yourself of this, the first step is see why the above description of the behaviour of the air molecules from t to $t+\Delta t$ must be nomically possible, leaving past history out of account. This follows from the fact that all our deterministic dynamical theories admit a "time-reversal" operation τ on the state space such that whenever $q=\Phi(x,p)$, $\tau(p)=\Phi(x,\tau(q))$. In the case of gas molecules, τ reverses all the velocities but leaves all positions alone: so if p is a point at which the molecules are all in the corner, $\tau(p)$ is too. τ also preserves the property of being such that the air is in thermodynamical equilibrium. So consider some state p at which the air is in the corner such that at $\Phi(\Delta t,p)$ the air is in equilibrium. Then $q=\tau(\Phi(\Delta t,p))$ is an air-at-equilibrium state with the surprising feature that $\Phi(\Delta t,q) = \tau(p)$ is an air-in-the-corner state. Moreover, since τ is volume-preserving, and since the set of air-in-the-corner states which evolve after Δt into air-at-equilibrium states has nonzero volume, the set of air-at-equilibrium states which evolve after Δt into air-in-the-corner states must also have nonzero volume.

What we still need to do is to convince ourselves that some of these strange states lie at the ends of trajectories approximately matching the pre-t history of the actual world.[20] One way of doing this is to appeal to what I will dub the 'Independence Conjecture', a certain plausible mathematical claim about the behaviour of dynamical maps that plays a central role in statistical mechanics. This conjecture is expressed in terms of the notion of a "macrostate": a set of dynamical states that agree on a certain set of statistical quantities, such as the mean temperature, pressure, density, and momentum of gas within each cell in some fine-grained lattice. The conjecture is that "the macropresent screens off the macrofuture from the macropast": in the probability distribution that we get by restricting the natural volume measure to a particular macrostate, facts

about *future* macrostates are, approximately, probabilistically independent of facts about *past* macrostates.[21] If this is true, then the proportion of points that evolve after Δt into air-in-the-corner states within the current macrostate approximately equals the proportion of such points within the much smaller set of points whose evolutions into the past correspond to the sequence of macrostates that have actually been instantiated. This provides one good sense of "approximate match" in which we can be confident that some of the air-in-the-corner worlds approximately match actual history.

The Independence Conjecture is not the sort of thing for which we should expect a proof any time soon even for a toy model system, but there are several reasons to regard it as plausible. Investigations of chaotic systems help to boost its *a priori* plausibility, since it has turned out to be common for such systems to be well-modelled as random walks under various kinds of coarse-graining. There is also considerable empirical support for the Independence Conjecture. As Wallace (forthcoming) persuasively shows, it—or rather, a more general conjecture of which it is a special case—plays a ubiquitous role in statistical-mechanical "derivations" of equations governing macro-quantities (such as the Boltzmann equation); and many of these equations have proved to be fantastically accurate.[22] And remember too that the claims of probabilistic independence yielded by the conjecture are much stronger than the mere existence claims we are concerned with—for example, the claim that there is a nomically possible world whose history approximately matches actuality until t at which the air subsequently squeezes itself into the corner.

If you are convinced that there are worlds like that, you should be willing to take the same view about many other "thermodynamical miracles". For example, you should think that for any given direction, it is nomically consistent with the approximate course of history up to now that the air molecules around you are about to move around so as

to impress a substantial net force on you in that direction: you could be whisked away by a sudden, anti-thermodynamic gust of wind. Moreover, given the Independence Conjecture, the occurrence of one thermodynamic miracle does not affect the probability of subsequent such miracles. So we should be able to add more such gusts of wind to keep you going in the relevant direction, and eventually set you gently on your feet at any given point on the surface of the Earth. For similar reasons, we can be confident that among the nomically possible worlds which match actuality macroscopically up to now, there are some where a succession of anti-thermodynamic puffs of wind in the vicinity of my computer keyboard types out the complete text of *Hamlet*.[23]

I am not suggesting that we will often want to appeal to thermodynamic miracles in developing ordinary counterfactual suppositions. Indeed, appealing to such miracles is often a bad idea even when it allows for a longer interval of approximate match with actuality. Suppose that, on the phone to Mary at t, Fred speaks the truth by saying "If I were there right now, I would give you a hug". On the operative interpretation of the counterfactual, how do we think Fred would have got to be with Mary at t? Would he have been whisked there quickly by a recent, anti-thermodynamic puff of wind, or would he have got there by a less showy method requiring a somewhat earlier divergence from the approximate course of actual history? The latter option seems better. If we choose the puff of wind, we will need to combine it, rather artificially, with further unusual goings-on in Fred's brain to ensure that he arrives still in a mood to give Mary a hug, rather than utterly befuddled or panic-stricken by his adventure.[24] And—more decisively—appeals to thermodynamic miracles will lead us badly astray in evaluating arguments like 'Nothing mysterious has ever happened to Fred, Mary, or Jane; so if they were all here right now, the people who have experienced mysterious events would be a minority'. I have dwelt on the nomic possibility of anti-thermodynamic goings-on as a

way of providing a concrete grasp on the surprisingly weak extent to which the approximate truth about history up to t constraints the course of history after t. Once we have appreciated that the set of nomically possible worlds approximately like actuality up to t contains even these bizarre worlds, we should be much more confident that it also contains worlds where the antecedents of the counterfactuals we are interested in come true in unremarkable ways (so long as those antecedents concern times long enough after t for the mere fact of such a transition not to be in itself remarkable).

This reassurance is strongly confirmed when we turn from gases in boxes to models of the kinds of processes that we treat as random in our successful high-level theories—rolling dice, spinning roulette wheels, cells undergoing meiosis, spheres of enriched uranium on the verge of going critical.… When such systems are analysed within classical dynamics, one generally finds the selection between the various outcomes to which our high-level theories assign positive chances to be determined by initial conditions in a very sensitive way.[25] Within the relevant part of the state space, any small sphere whose volume exceeds some microscopic threshold will contain points that evolve into each of the possible outcomes.[26] And given the Independence Conjecture, we can expect the same to be true of the sets corresponding to worlds that have stayed close to actuality throughout the past, even though these sets will not look like spheres unless we blur our eyes to their filigreed microscopic structure. Thus, for example, we have strong reason to believe that whenever a roulette wheel is spun, the set of nomically possible worlds that stay very close to actuality until the spin contains worlds in which the ball ends up on each number on the wheel. Similarly, given that the firing of a neuron receiving inputs at a given rate is successfully modelled as a chancy process, we have reason to believe that any medically possible pattern of firings of someone's neurons can be realised in the first few seconds after t in a nomically possible world that ap-

proximately matches actuality up to t. For most everyday counterfactuals, divergences of this sort will give us exactly what we are looking for.

There are exceptions: although fine-grained dependence on initial conditions is pervasive, it is not omnipresent. Sometimes, there will just be no way to make the antecedent of a counterfactual about a particular time true in a nomologically possible world without allowing that world to diverge substantially from actuality at much earlier times. One might worry that in such cases, allowing the past to vary to the required extent will disrupt the assignments of truth-values which we find intuitive in "standard" contexts. Consider

(6) If a big comet had hit Washington, D.C. yesterday afternoon, the U.S.A. would have been left without a President.

(6) seems warranted if we know that everyone in the line of succession was in D.C. yesterday afternoon. But the antecedent of (6) is probably not true in any nomically possible world that approximately matches actuality until yesterday. Although there are nomically possible worlds where comets suddenly accelerate due to thermodynamically miraculous ejections of matter, there is a limit to the velocities that can be achieved in this way without the comet's ceasing to exist. And given how far we are from all the big comets, this means we will need quite a long interval before the counterfactual impact—weeks? years?—during which the trajectory of the comet is very far from its actual trajectory. As one reflects on this, it is tempting to start entertaining "backtracking" thoughts like 'If a comet had hit D.C. yesterday afternoon, astronomers would probably have seen it coming at least a day earlier and ordered an evacuation, so that the U.S.A. would still have had a President'. One live option for proponents of *Laws* is to endorse these second thoughts, and say that (6) is false even by the standards of its original con-

text. As error theories go, this is relatively moderate: if the heuristics we use in evaluating counterfactuals are reliable for the most part, it is easy to see how they might lead us astray in certain unusual cases such as those involving heavenly bodies. However, this kind of error theory is certainly not inevitable. Although in evaluating (6) we cannot simultaneously hold fixed the laws and the entire approximate truth about history before yesterday afternoon, there is no obvious reason why we should not be able to hold fixed the laws and the approximate history of the *Earth* before yesterday afternoon, including the facts about the observations made by astronomers as well as the locations of politicians. Given the pervasiveness of dynamical chaos in the Earth's atmosphere and in people's brains, there certainly are nomically possible worlds where unlikely events on some comet first set it on a collision course for Washington, and then a series of further unlikely events on Earth prevent the comet's approach from having any of the expected macroscopic effects on us. For example, we can cook up unusual arrangements of air molecules between the astronomers' telescopes and the comet which deflect the incoming photons from the comet whenever the astronomers are looking, or unusual patterns of firing in the astronomers' brains which prevent them from noticing the comet even when the photons are streaming into their eyes. The fact that *most* such arrangements would blow up quickly into macroscopic differences is not a reason to doubt that *some* such arrangements allow us to preserve the historical facts that seem to be contextually held fixed in the face of the comet's approach.

Another worry concerns counterfactuals like (7), which will be credible only if we are allowed to hold fixed certain *microscopic* facts:

(7) If we had aimed the electron microscope a tenth of a degree further to the left, the image of that gold atom would have appeared in the centre of the screen.

One might worry that (7) will not be well-supported if we can only hold fixed the approximate course of history. But this worry assumes that the relevant notion of "approximate match" always has to be understood in terms of something like the standard distance measure on state space or the standard partition into macrostates, in which all particles are treated on a par. In fact, however, there is room for plenty of contextual flexibility as regards what exactly gets fixed. We can even say (if we want to) that in a context where certain specific atoms and times are salient, the facts to be held fixed include the *exact* positions of those atoms at those times. Although such constraints will slightly reduce the dimensionality of the set of states whose histories count as "approximately matching" that of the actual world, there is still every reason to expect that some of these states will correspond to worlds where the antecedents we are concerned with are true.[27]

Summing up: even if we accept *Laws*, we can and should think that if a normal person in a deterministic world had blinked at t, the course of history until shortly before t would have been different only in negligible microscopic respects. The option of denying *Past* while keeping *Laws* is thus far less radically revisionary that one might initially have supposed.[28]

5. Denying *Past*: does it lead to backwards causation?

There is a very different argument for *Past* which turns on the plausible premise that deterministic worlds need not contain backwards causation. Upholding *Laws* requires saying that if determinism is true, then if a given normal person had blinked at t, the course of history would have been different all along (albeit only microscopically so). According to the objection, even this amount of counterfactual dependence of earlier history on later history entails that there is backwards causation: some later events

cause earlier ones, or some facts about earlier times hold because of facts at later times in the causal sense of 'because'.[29]

It is important to note that Lewis, the most prominent proponent of counterfactual miracles, is in no position to make this objection. According to Lewis, if the world is deterministic and *x* didn't blink at *t*, then if *x* had blinked at *t*, *part* of the earlier history of the world would have been different. He allows for a short transition period—what Bennett (2003, §81) calls a 'ramp'—such that if a given proposition about a time *t* had been true, the miraculous divergence from actual history would have taken place somewhat earlier than *t*. And there is nothing independently plausible about the thesis that *short-range* counterfactual dependence of facts about earlier times on facts about later times need not involve backwards causation, whereas *long-range* counterfactual dependence can only obtain in worlds featuring backwards causation.[30]

The only view about counterfactuals under determinism that would allow someone to make the backwards-causation argument without being vulnerable to a *tu quoque* is a "big miracle" view that holds fixed the exact course of history *in its entirety*: when the antecedent of a counterfactual is about events at and after *t*, and its consequent is entirely about history before *t*, the counterfactual automatically gets the same truth value as its consequent. So for example, given that I was in fact in my kitchen yesterday from 11am to 1pm, if I had begun driving to Princeton at noon, I would still have been in my kitchen at every time earlier than noon. In other words, I would have teleported instantaneously into the car from the kitchen.[31]

The "big miracle" view faces an obvious objection. If it is correct, shouldn't we be drawing on *all* truths about times earlier than *t* in reasoning about counterfactuals whose antecedents are about times after *t*, and thus taking seriously counterfactuals like 'If I had begun driving to Princeton at noon yesterday, I would probably have crashed

because of my extreme surprise at my sudden teleportation'? Such reasoning would make trouble for many everyday counterfactuals.[32] However, if we were willing to embrace big counterfactual miracles, we could get around this worry by adopting a "huge miracle" view, in which the big miracle required for the truth of the antecedent would have been accompanied by a range of ancillary miracles which prevent the sudden changes involved in the big miracle from having any of their expected effects. For example, we could say if I had begun driving to Princeton at noon yesterday, not only would I have miraculously teleported into the car, but further miracles would have occurred to prevent me from being surprised by the transition, to fill the vacuum I left behind in the kitchen, etc.[33]

However, besides straining credibility, the "huge miracle" view is of no use for the purpose of preserving the thesis that counterfactual dependence suffices for causation. The problem is that the ancillary miracles will induce many cases of counterfactual dependence without causation. Suppose that someone was making a video in my kitchen at noon. Then the "huge miracle" view will presumably say that if I had started driving to Princeton at noon, the portion of the recording made at 11.59 would have changed at noon in such a way as to show the kitchen as empty at 11.59. If counterfactual dependence suffices for causation, this entails that my failure to begin driving to Princeton at noon *causes* the later presence on the camera of a recording showing me sitting in the kitchen at 11.59. But in fact there is no direct casual relation between these events: they are merely effects of a common cause (my earlier decision not to go to Princeton).

The thesis that counterfactual dependence (understood according to ordinary standards) suffices for causation is simply not well-supported enough to be worth saving by such heroic means. So whether or not we accept *Laws*, we should allow at least for short-range future-to-past counterfactual dependence, and refrain from endorsing any

account of causation on which counterfactual dependence is sufficient for causation.[34] This leaves the coast clear for those of us who do accept *Laws* to say the very same thing about the long-range future-to-past counterfactual dependence that our view requires. Indeed, insofar as this long-range dependence involves only microscopic facts at the earlier times, it is considerably *less* worrying from the point of view of the theory of causation than the short-range dependence that everyone needs. For on many accounts, causation, or the temporal directionality of causation, is in some sense founded in certain patterns that obtain in the macroscopic world, so that we should be cautious whenever we attempt to extend our thinking about causation into the microscopic realm.[35]

6. Denying *Laws*: bizarre counterfactuals

So far, I have argued against *Triviality*; argued for *Closure*; and rebutted some arguments for *Past*. In this and the following section I will make a positive case for *Laws* (and thus against *Past*).

Suppose that L is a simple, true, deterministic law, and that Frank, a philosopher of physics, has devoted his career to defending the truth of L.[36] He is having a public debate with Nancy, who maintains (wrongly) that there are isolated exceptions to certain generalisations that follow from L, so that L is false. If we keep *Past* and *Closure*, we have to say that if the circumstances of the debate had been different in any way whatsoever—for example, if someone had put a glass of water on Frank's lectern, or rudely interrupted his talk—then Nancy would have been right and Frank wrong. Thus (8) and (9) are true:

(8) If we had given Frank a glass of water, his whole career would have been devoted to a mistake.

(9) If you had told Frank that his whole career was devoted to a mistake, you would have been right.

These are bizarre and unpalatable consequences.

Of course *Laws* itself is intuitively plausible, so it is not surprising to find that denying it leads to counterintuitive results. But the case does show that the revisionary consequences of preserving *Past* are more far-reaching than its proponents might have hoped. They are not limited to claims of the sort that only philosophers and scientists ever discuss, such as the negation of *Laws*; they include claims of a more quotidian, practical cast, like the negations of (8) and (9).

Could the problem be dissolved by invoking context-sensitivity? The thought would be that (8) and (9) elicit an unusual interpretation of the conditional under which the truth of L is held fixed, although it is not held fixed in ordinary contexts. But it is hard to see how such shiftiness could emerge from any credible account of what drives the resolution of context-sensitivity. (Notice that the judgment that (8) and (9) are false is, if anything, even clearer if we imagine that the speakers have no idea what Frank's career has been about.) Moreover, the relevant judgments seem to integrate seamlessly into the ordinary pattern of reasoning characterised by *Holding History Fixed*. For example, if we think that Frank is right and Nancy wrong, we may draw on our memory of past debates to justify (10):

(10) If we had given Frank a glass of water, that would have been the fifth debate in a row in which the defender of the true view was the only one with a glass of water.

This makes it hard to believe that the context elicited by our examples is very different from standard contexts with respect to what gets held fixed.

It is plausible that in judging (8) and (9) false, we are in some sense implicitly relying on the principle that laws can be held fixed in evaluating counterfactuals with nomically possible antecedents. If so, arguments based on these judgments will be dialectically ineffective against someone who has wholeheartedly internalised the commitments of the *Laws*-denying strategy, which include the denial of that principle. But this is not the dialectical context in which I am giving this argument. The argument is aimed at those who, like myself, feel the pull of all the premises of the CCA and of the negation of its conclusion, and are trying to weigh the costs of the various ways out of the paradox. In this setting, it is not a problem if our resistance to (8) and (9) is caused by our deep-seated attachment to some general principle that entails *Laws*: the argument can still show that the costs of denying *Laws* are higher than one might otherwise have thought.

What *would* make the argument dialectically ineffective would be the existence of a parallel and equally compelling argument for *Past*. If such an argument could be given, we would have made no progress towards a reasoned way out of the paradox. So, let's consider how it would need to go. We would need to imagine a character whose career has been devoted to some proposition H about the intrinsic character of some initial segment of history, such that the conjunction of H with the laws necessitates some everyday truth, for example that there is no water glass on a certain lectern during a certain debate. But if H is supposed to be expressed in the language of physics, a story like this cannot be realised in any remotely realistic possible world. Any proposition about the early universe which nomically necessitates the absence of a water glass must be sensitive to the value of a huge number of parameters (such as the distances between particles) to some preposterous number of decimal places. Ordinary human beings cannot even formulate such propositions, much less devote their careers to them.[37] The

best we can do is to get at the relevant intrinsic properties indirectly, by saying things like 'The complete intrinsic truth about the first million-year segment of history is such that, in conjunction with L, it entails that there is no glass of water on this lectern during this debate'. We can, indeed, imagine a character—call her Lavinia—who has devoted a career to this claim: this requires a ridiculous set of priorities, but not superhuman cognitive powers. Proponents of *Laws* will then have to accept (11):

(11) If we had put a glass of water on her lectern, Lavinia's career would have been devoted to a mistake.

But (11) is clearly much less repugnant than (8). Given Lavinia's bizarre choice to stake her career on a claim specified by reference to this particular lectern and time, it is really not so odd to think that she thereby made the success of her career a hostage to our choices about the distribution of water glasses.

Our argument thus picks up on an important contrast between *Past* and *Laws*. If determinism is true, then truths about earlier times that nomically necessitate facts about what people do at later times are extremely complicated when specified in physical terms—far too complex for anyone to entertain, at least under the modes of presentation associated with such terms. By contrast, if there are deterministic laws of nature, they are probably quite simple. The contrast between the Frank and Lavinia examples shows how this asymmetry can be put to work to break the dialectical stalemate between *Past* and *Laws*.[38]

One might worry that the argument would be undermined by the fact that the relevant asymmetry is metaphysically contingent. The initial state of the universe could have been much simpler than it actually was—perhaps even as simple as the laws of nature actually are. Meanwhile, assuming that the laws could have been different at all,

they could have been at least *somewhat* more complex than they actually are. It is not obvious how much more complex they could be: if a Humean "best system" account of lawhood like that of Lewis (1994) is true, it may be impossible for them to be *vastly* more complex. But I don't want to assume Humeanism, and there is no obvious reason for anti-Humeans to place any limits on the possible complexity of laws. Indeed, given anti-Humeanism, there may even be metaphysically possible, deterministic worlds where the asymmetry is completely reversed: the laws are as complex as the actual initial conditions, while the initial conditions are as simple as the actual laws.

If there are possible worlds like this, then probably, at some of them, there are people who devote their careers to defending some simple truth about the initial conditions which, in conjunction with the (ultra-complex) laws, entails that there is no glass of water on their lectern. Since holding fixed the laws in this scenario will entail counterfactuals just as intuitively repugnant as (8) and (9), one might conclude that the mode of reasoning that leads us to reject such counterfactuals ultimately does nothing to settle the question whether it is the laws or the initial conditions that should be held fixed.

It is not obvious that this conclusion would be warranted. Given that the actual laws are simple, one might reasonably insist that judgments about worlds that resemble actuality in this respect deserve more weight than judgments about bizarre worlds with ultra-complex laws. But even if we regard the judgments as being dialectically on a par, the conclusion neglects the possibility that we might regard the question whether the laws or the initial conditions are counterfactually robust as contingent, holding that the laws are robust at possible worlds where the initial conditions are much more complex than the laws, while the initial conditions are robust at worlds where the laws are much more complex than the initial conditions. This is in fact the option I would favour, assuming that there are any possible worlds of the latter sort. However, since the laws of

such worlds are so very dissimilar from the deterministic laws that are actually serious epistemic possibilities, it will be convenient if we avoid the need to take account of them in discussing the CCA, by stipulatively interpreting the notion of "normality" in *Past* and *Laws* in such a way that people in deterministic worlds with ultra-complex laws do not count as "normal". On this interpretation, considerations involving such worlds will not threaten *Laws*.

What about possible worlds where the past and the laws are *both* simple? If there are worlds like this where there are people capable of grasping the simple propositions in question, then there will be no escaping the conclusion that counterfactuals just as bizarre as (8) and (9) are true in those worlds. Someone might argue that since the mode of reasoning that prompts us to reject (8) and (9) leads us astray when we are dealing with worlds where the laws and the past are both simple, this mode of reasoning should not be trusted even when applied to realistic worlds where the past is much more complex than the laws. But this is not a very impressive argument: the fact that a certain mode of reasoning goes wrong in some very esoteric possible cases does little to undermine its force when applied to more realistic cases. And in any case, it is doubtful that there could even *be* any people in a deterministic world where the laws and past were both simple. After all, people are necessarily rather complex things, and it is hard to see how the required complexity could be generated from the evolution of simple states in accordance with simple deterministic laws.

(Remember that we are not here concerned merely with simplicity in the *qualitative* profile of the initial state. This is, plausibly, compatible with the existence of people in great abundance, even given simple deterministic laws. For example, imagine a world consisting of a plenitude of non-interacting Newtonian universes, one for every geometrically possible initial arrangement of finitely many particles.[39] But the everyday

propositions that occur in problematic counterfactuals like (8) and (9) are non-qualitative: they concern particular people, lecterns, and so forth. For these haecceitistic propositions to be metaphysically necessitated by the complete qualitative characterisation of the populous worlds in question, the relevant objects would have to have extremely demanding essences, such that it would be metaphysically impossible for them to instantiate any of the qualitative profiles that other objects actually instantiate. This is far-fetched.[40] Assuming that people at the populous worlds don't have such demanding essences, the only propositions about the past that nomically necessitate the relevant everyday propositions will be massively haecceitistic—for example, propositions which fully specify the distribution of physical properties and relations over some enormous collection of particular particles. It is doubtful whether anyone could possibly even grasp a proposition like this, let alone devote a career to it.)

The metaphysical contingency of the fact that the past is more complex than the laws thus fails to undermine the argument from the falsity of (8) and (9) to the truth of *Laws*.

7. Denying *Laws*: emotion and decision

The previous section's argument for *Laws* turned on some vivid pre-theoretical judgments about the falsity of certain counterfactuals. The present section will bolster those judgments by drawing out certain unpalatable consequences which would follow from the problematic counterfactuals in conjunction with certain plausible claims about the role of counterfactuals in our emotional and practical lives.[41]

First: certain emotions are intimately bound up with counterfactual thought. Regret is the canonical example. Regretting doing something seems to require wishing that one had not done it; and it seems irrational to wish one had not done something while believing that things would have been worse if one had not done it, in the respects one cares about. When our mistakes turn out to have been for the best, we may still feel

ashamed of them, but we can no longer unequivocally regret them. For example, suppose Mary thoughtlessly left her passport at home while on her way to a conference, and so had to cancel her talk, inconveniencing the conference organisers. But while waiting around the airport, Mary met Fred, the love of her life. If she hadn't forgotten her passport that day, she would never have met him. Knowing this, Mary should be *glad* that she forgot her passport, rather than regretful.[42]

Now consider our character Frank from section 6. Suppose that he greatly values not having devoted his career to a mistake. If regret is constrained by beliefs about counterfactuals in the way I have suggested, then if Frank comes to believe that his career would have been devoted to a mistake if he had not acted in a certain way, he will be unable to rationally, unequivocally regret acting in that way (assuming nothing else of comparable importance is at stake). In that case, coming to believe *Past* will give Frank a sovereign remedy against regret! Whenever he does anything, no matter how foolish, he can immediately afterwards look back and think, 'If I hadn't done that thing, my whole career would have been devoted to a mistake!'. If he really believes this, and no other comparably weighty values are in play, he should be, all things considered, *glad* that he did the foolish thing. But it seems crazy to suppose that coming to believe the truth about how counterfactuals work under determinism would put Frank in this situation.[43]

The second group of problems arise if beliefs about counterfactuals play a central role in *rational decision*, as they do on what is arguably the most widely-accepted form of decision theory, the so-called "causal decision theory" sketched by Stalnaker (1981 [1972]) and developed by Gibbard and Harper (1978). According to this view, ideal rationality involves a disposition to choose only actions whose *expected counterfactual utility* is at least as high as that of any other available action, where (roughly speaking) the

expected counterfactual utility of an action is one's estimate of how well things would go, in the respects one cares about, if one performed it.

Suppose that Frank is faced with some decision, is highly confident that L is true and that the distant past would be exactly the same no matter what he did, and cares much more about not devoting his career to a mistake than about any other relevant value. Then the action with the highest counterfactual expected utility for Frank will simply be whichever action he is most confident he will in fact perform. The more confident he is that he is going to do a certain thing, the more confident he will be that if he were to do anything else, his career would have been devoted to a mistake. Imagine that he has a track record of acting in some intuitively foolish way—for example, by buying ten lottery tickets every time he leaves his house. Then one way for him to satisfy the demands of causal decision theory when he next goes out is to buy the tickets again, having predicted on inductive grounds that he will buy them, and thereby become confident that if he weren't to buy them, his career would have been devoted to a mistake.

Of course, if for some reason Frank becomes confident, despite his track record, that he will *not* buy lottery tickets this time, not buying the tickets will become the option maximising expected counterfactual utility. So if his actions fit his credences and preferences in the way advocated by causal decision theory, this prediction will also be self-fulfilling. Frank's situation is like that of someone who is convinced that a reliable predictor has predicted their behaviour and then undertaken to reward them if the prediction comes true. Whatever we think rationality requires in such situations of "rational feedback" (Reaber 2012), it is implausible that believing the truth about how counterfactuals work under determinism would put Frank in such a situation.[44]

The problem is even sharper when we consider negative rather than positive feedback. Suppose that Frank's arguments are so compelling that Nancy changes her mind

and becomes convinced that L is true. She is sad about this, since she wishes that she had not devoted her career to a mistake. But even though she is now confident that she *did* devote her career to a mistake, her strong preference not to have done so will play out bizarrely in every decision situation, if she believes that the distant past is counterfactually independent of her actions and is disposed to maximise expected counterfactual utility. In general, the option with highest expected counterfactual utility for Nancy will be the one she is *least* confident she is going to choose. If she is even moderately good at introspection, so that she has a roughly accurate sense of her credences, her preferences, and her disposition to maximise expected counterfactual utility, her practical life will be paralysed. She will never be confident that she is going to choose X over Y, since once she becomes aware of this confidence, she will expect that it will tend to lead to her choose Y over X. The only equilibrium states for her will be states in which she is no more confident that she will choose one option rather than any of the others; and she will be stuck in this state of uncertainty until she eventually surprises herself by acting.

Ahmed (2013, 2014) presents a somewhat similar case, called *Betting on the Laws*, which can also serve as an argument against the combination of counterfactual decision theory with *Past*. Ahmed's protagonist, Alice, is very confident on reasonable grounds that a certain proposition L* is a true, deterministic, law. She must choose between affirming L* and denying it (affirming its negation); and cares only about affirming the truth on this occasion. Assigning value 1 to speaking truly and value 0 to speaking falsely, Alice's expected counterfactual utility for affirming L* equals her credence that L* would be true if she affirmed it, while her expected counterfactual utility for denying L* equals her credence that L* would be false if she denied it. But suppose that Alice is certain that that there is some history-proposition which would be true no matter what

she did now, and that L* is deterministic. Then, conditional on the hypothesis that L* would be true if she affirmed it, Alice will be certain that if she had affirmed L*, the conjunction of L* with any true history-proposition would have entailed that she affirms L*. Since she is certain that at least one history-proposition which would be true if she affirmed L* would also be true if she denied L*, she must therefore be certain, conditional on the hypothesis that L* would be true if she affirmed it, that L* would be false if she denied it. Alice's credence that L* would be false if she denied it must therefore be greater than or equal to her credence that L* would be true if she affirmed it. In fact it must be strictly greater, so long as she assigns nonzero credence to the proposition that L* would be false either way. Causal decision theory thus entails that Alice is rationally required to deny L*, despite her high credence in L* and her complete devotion to the truth. But it is absurd to suppose that being certain of the true view of counterfactuals under under determinism could require such a bizarre mismatch between Alice's beliefs and preferences and her verbal behaviour.[45]

Note that Alice's practical situation does not involve the kind of rational feedback that characterised Frank and Nancy's situations in the earlier examples. We can imagine Alice anticipating the reasoning of the previous paragraph, and becoming confident that she will deny L*. Given her high credence in L*, she is disappointed, since she now is confident that she will speak falsely. Nevertheless, she still regards denying L* as the lesser evil: although her credence that she would speak truly if she denied L* is low, her credence that she would speak truly if she affirmed L* is even lower.

These arguments for *Laws* depend of course on a controversial view about decision theory, namely causal decision theory—or to be precise, Stalnaker-Gibbard-Harper counterfactual decision theory.[46] This not the place to rehash the well-known arguments for the superiority of causal decision theory over its principal rival, evidential de-

cision theory. However, since Ahmed uses *Betting on the Laws* to argue for evidential decision theory and against causal decision theory, it is worth considering his argument to make sure that it does not undermine mine.[47] It can be reconstructed as follows. (i) Any theory that captures a certain intuitive thought about the relevance of causation to decision—namely, 'that you should realise the option that you expect causally to make you best off, regardless of what doing so might reveal about any of its non-effects that might also matter to you but which are beyond your present influence'—wrongly entails that Alice should deny L*.[48] (ii) Any theory other than evidential decision theory that does not capture that intuitive thought is 'unmotivated' and 'arbitrary'.[49] So (iii) we should accept evidential decision theory.

I am not convinced by either of Ahmed's premises, but I will focus here on (ii).[50] Even if Ahmed is right to think that counterfactual decision theory fails to capture the relevant intuitive thought about causation when it is combined with the denial of *Past*, it seems to me that it is still well-motivated, both by judgments about what people should do in certain particular cases—especially Newcomb's problem and the "medical Newcomb cases"—and by a compelling, intuitive account of why those are the right judgments. After all, in justifying the decision to take two boxes in Newcomb's problem, it is extremely natural to reach for counterfactual language: 'Whatever is in the opaque box, I know I would get a thousand dollars more if I were to take both boxes than I would if I were to take only one'. If we are challenged to *justify* these counterfactual claims, it is indeed tempting to appeal to further premises about causation. But since it is a hard task to explain exactly what counterfactuals have to do with causation (see §5), I see nothing *ad hoc* or unstable about a view that implements the intuitive thought about the relevance of counterfactuals in the way that Stalnaker-Gibbard-Harp-

er decision theory does, while admitting that causal independence is not always sufficient for counterfactual independence.[51]

The combination of *Betting on the Laws* with the standard arguments against evidential decision theory thus provides robust support for a package consisting of a counterfactual decision theory like Stalnaker's, Gibbard and Harper's (or Lewis's), combined with a law-preserving account of counterfactuals under determinism like the one I have been arguing for.[52]

8. Lessons for other consequence arguments?

The CCA belongs to a family of structurally isomorphic arguments, some other members of which have been widely discussed in the literature on free will. In this concluding section, I will briefly consider the question how the foregoing reflections on the CCA might bear on our evaluation of other arguments in this family.

The common structure of these arguments can be presented in the form of a valid argument-schema:

Schematic Consequence Argument (SCA):

Past Necessarily, whenever x is normal at t, there is a true history-proposition p such that $R(x, t, p)$.

Laws Necessarily, whenever x is normal at t and p is a true law of nature, $R(x, t, p)$.

Closure Necessarily, whenever x is normal at t and p is metaphysically necessitated by a set S of propositions such that $R(x, t, q)$ for every $q \in S$, $R(x, t, p)$.

Triviality Necessarily, whenever x is normal at t and determinism is true, every true proposition p is such that $R(x, t, p)$.

Here 'R(x, t, p)' stands for some relation between agents, times, and propositions. We can recover the CCA by taking 'R(x, t, p)' to be 'p would have been true if x had blinked at t'.

In the literature on free will, we find a bewildering variety of other candidates which we could substitute for 'R(x, t, p)' in the SCA to generate *prima facie* compelling arguments for conclusions with some claim to the name 'incompatibilism'. They include the following:

(i) p is true, and x has no choice at t about the fact that p is true (van Inwagen 1983: 93 ff.)

(ii) p is true, and at t it is not up to x whether p is true (van Inwagen 1983: 56)

(iii) p is true and at t, x cannot render p false (van Inwagen 1975)

(iv) At t, x cannot act in such a way that p would not be true (Fischer 1994)

(v) p is true and at t, x cannot make it the case that p is false (Ginet 1990)[53]

(vi) p is true and every maximally strong proposition whose truth x can ensure at t entails p (van Inwagen 2000)

(vii) p is true and nothing x can do at t is such that if x did it, this deed would be or cause an event whose occurrence is inconsistent with p (Lewis 1981a).

(viii) p is true and nothing x can do at t is such that if x did it, p would be false (Lewis 1981a; Fischer 1994: 8).

(ix) p is true and everything x can do at t is such that if x did it, p would be true.

While it is plausible that some of the expressions on this list are equivalent, there are no particular pairs whose equivalence is obvious.

Of these nine versions of the SCA, version (ix) is, I would argue, stronger than any of the others (except for any that are equivalent to it).[54] The reason turns on the plausibil-

ity of *Closure*. *Closure* in version (ix) is plausible for the same reason as *Closure* in the CCA: both follow from *Weak General Closure* (see section 3 above). By contrast, *Closure* in version (vii) is widely recognised to be false, since the occurrence of a certain event can be inconsistent with a proposition p without being inconsistent with either of two propositions which jointly entail p.[55] *Closure* in version (viii) is also highly controversial, since it should be rejected by those who reject the principle of Conditional Excluded Middle, according to which instances of the schema 'Either if P, it would be that Q or if P, it would be that not-Q' are true in every context.[56] Similar considerations apply to versions (i)–(vi). Insofar as there was reason to think that one of these versions was *not* equivalent to version (ix), there would also be reason to worry that *Closure* in that version might suffer from the same problems that beset *Closure* in versions (vii) and (viii).

Whether this comparative judgment is right or wrong, version (ix) of the SCA is a formidable argument, worth thinking carefully about. Let's spell it out explicitly:

Ability Consequence Argument (ACA)

Past Necessarily, whenever x is normal at t, there is a true history-proposition p such that p would still have been true if x had done any of the things x can do at t.

Laws Necessarily, whenever x is normal at t and p is a true law of nature, p would still have been true if x had done any of the things x can do at t.

Closure Necessarily, whenever x is normal at t and p is metaphysically necessitated by a set of propositions each of which would still have been true if x had done any of the things x can do at t, p would still have been true if x had done any of the things x can do at t.

Triviality Necessarily, if x is normal at t and determinism is true, every true proposition p is such that p would still have been true if x had done any of the things x can do at t.

It is just a small further step from *Triviality* to

Incompatibilism Necessarily, if x is normal at t and determinism is true, everything x can do at t is something that x in fact does.

To bridge this gap, all we need is the uncontroversial claim that necessarily, normal people cannot do things such that contradictions would have been true if they had done them.[57]

The ACA and arguments relevantly like it have convinced many people that *Incompatibilism* is true. This is not surprising, since the ACA is valid and all of its premises are pre-theoretically compelling, while its conclusion is not especially repugnant. By contrast, although the CCA has similarly plausible premises, its conclusion *is* repugnant: I know of no-one who has embraced it, despite the fact that the basic problem has been known for decades. Those who reject the conclusion of the CCA will need to reject one of its premises. But this matters to our evaluation of the ACA, since the *sources* of the intuitive plausibility of the premises of the CCA and those of the ACA seem quite similar. As noted above, both versions of *Closure* follow from *Weak General Closure*. Both versions of *Laws* are rooted in one of the central theoretical roles of the concept of a law of nature, the idea of laws as supporting counterfactuals.[58] And both versions of *Past* owe their basic appeal to a certain picture of what it means for the past to be "over and done with". Moreover, all the premises can be argued for abductively, on the grounds that

they explain and systematise a range of plausible judgments about particular cases.[59] This kinship between the two arguments creates some presumption that the corresponding premises should stand or fall together. Anyone who rejects some premise of the CCA while accepting its counterpart in the ACA thus faces the challenge of finding a principled way to drive a wedge between the two. It is no longer dialectically sufficient simply to appeal to the pre-theoretical plausibility of the relevant premise of the ACA, or to an abductive inference based on the pre-theoretical plausibility of certain of its consequences. The failure of the CCA-premise tends to undermine our justification for trusting the relevant pre-theoretical judgments, or for accepting the abductive inference.

The challenge seems especially difficult to answer if, as I have argued, the false premise in the CCA is *Past*. Those who reject *Laws* in the CCA but not in the ACA might, perhaps, hope to meet the challenge by arguing for some intimate link between the concept of ability and that of a law of nature. Both are associated with familiar modalities, and there may be some way to make a case that "practical" possibility entails nomic possibility that is neither undermined by coming to think of laws as counterfactually fragile, nor dependent on some completely different argument for incompatibilism. By contrast, while we can define a broad notion of historical possibility on which truths about the past, but not laws, count as historically necessary, this notion seems quite artificial and has no clear role in our ordinary thought. It is hard to imagine an argument that "practical" possibility entails this kind of historical possibility that would not be undermined by rejecting *Past* in the CCA and did not turn on the success of some independent argument for incompatibilism.

For those who do not feel the force of the presumption that the corresponding premises of the CCA and the ACA stand or fall together, it may help to consider the role the restriction to "normal" people plays in making the premises plausible. Dropping

this restriction sharply diminishes the appeal of *Past* and *Laws* in both arguments. We can understand stories about extraordinary people, such as gods and time travellers, who are able to blink at a certain time, and are such that if they had blinked at that time the laws or the distant past would have been different. It is not at all obvious that such stories are metaphysically impossible. But the notion of normality we have been using to set such cases aside is little more than a promissory note: we should always have expected to have to do some work in settling where to draw the relevant line. Once we have rejected *Past* or *Laws* in the CCA, we now think that one of the theoretically important classifications in the vicinity groups otherwise unremarkable people in deterministic worlds together with gods and time travellers. Given this surprising conclusion, we would surely need some special reason for insisting that the corresponding premise of the ACA is true on an interpretation of 'normal' that includes people in deterministic worlds.

The view I have been presenting about the dialectical significance of the puzzle of counterfactuals under determinism for the standard consequence arguments has not been widespread in the literature on incompatibilism.[60] One reason for this is the context-sensitivity of counterfactuals, which makes it hard to distinguish the surprising claims we are forced to accept in order to avoid the conclusion of the CCA from other, much less interesting claims. Recall the following sentences from §1:

(2) If determinism is true and x does not blink at t, then if x had blinked at t, that would have to have been because of a prior history of determining factors differing all the way back.

(2′) If determinism is true and x does not blink at t, then if x had blinked at t, a miracle would have to have occurred.

Both sentences, considered in isolation, seem like mere platitudes. As observed in §1, the 'would have to' construction that features in both (2) and (2′) seems, in other cases, to play a role in signalling that some kind of "non-standard" interpretation is intended. It is thus plausible that (2) and (2′) both express truths on the interpretations they naturally evoke. These truths are not particularly surprising or counterintuitive (at least under these guises). Merely acknowledging the unsurprising truth conveyed by (2) does little to undermine the intuitive appeal of *Past* in the ACA; likewise, acknowledging the unsurprising truth conveyed by (2′) does little to undermine the appeal of *Laws* in the ACA. Those of us who have decided whether to reject *Past* or *Laws* in the CCA will be tempted to use sentences like (2) and (2′) to make our choice seem obviously correct.[61] But we should resist this temptation. For the CCA presents a challenging paradox, any resolution of which will include a diagnosis of some tendency towards error latent in our ordinary methods of evaluating counterfactuals. The context-sensitivity of counterfactuals means that we will have to struggle to avoid confusing the revisionary claims involved in resolving the paradox with obvious truths. But, given the potential significance for other debates of the realisation that common sense is apt to lead us astray in this domain, the effort is worth making.[62]

References

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

—. 2014. *Evidence, Decision, and Causality.* Cambridge: Cambridge University Press.

Albert, David. 2000. *Time and Chance*. Cambridge, MA: Harvard University Press.

—. 2015. *After Physics*. Cambridge, MA: Harvard University Press.

Arntzenius, Frank. 2008. "No Regrets, Or: Edith Piaf Revamps Decision Theory." *Erkenntnis* 68: 277–97.

Beebee, Helen. 2003. "Local Miracle Compatibilism." *Noûs* 37: 258–77.

— and Alfred Mele. 2002. "Humean Compatibilism." *Mind* 111: 201–23.

Bennett, Jonathan. 1984. "Counterfactuals and Temporal Direction." *Philosophical Review* 93: 57–91.

—. 2003. *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

Dorr, Cian. n.d. "Finding Ordinary Objects at Some Quantum Worlds." Unpublished manuscript.

— and John Hawthorne. 2014. "Semantic Plasticity and Speech Reports." *Philosophical Review* 123 (2014): 281–338.

Field, Hartry. 2003. "Causation in a Physical World." In *The Oxford Handbook of Metaphysics*, ed. Michael J. Loux and Dean W. Zimmerman, 435–60. Oxford: Oxford University Press.

Fine, Kit. 2012. "Counterfactuals Without Possible Worlds." *Journal of Philosophy* 109: 221–46.

Fischer, John Martin. 1994. *The Metaphysics of Free Will.* Oxford: Blackwell.

Gibbard, Allan and William Harper. 1978. "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory*, ed. A. Hooker, J. J. Leach & E. F. McClennen, 125–62. Dordrecht: D. Reidel.

Ginet, Carl. 1966. "Might We Have No Choice?" In *Freedom and Determinism*, ed. Keith Lehrer, 87–104. New York: Random House.

—. 1990. *On Action*. Cambridge: Cambridge University Press.

Glynn, Luke. 2013. "Of Miracles and Interventions." *Erkenntnis* 78: 43–64.

Goggans, Philip. 1992. "Do the Closest Counterfactual Worlds Contain Miracles?" *Pacific Philosophical Quarterly* 73: 137–49.

Goodman, Jeremy. 2015. "Knowledge, Counterfactuals, and Determinism." *Philosophical Studies* 172: 2275–8.

Halpin, John F. 1991. "The Miraculous Conception of Counterfactuals." *Philosophical Studies* 63: 271–90.

Hawthorne, John. 2006. "Determinism *De Re*." In *Metaphysical Essays*, 239–43. Oxford: Oxford University Press.

Hitchcock, Christopher. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy* 98: 273–99.

Holliday, Wesley. 2012. "Freedom and the Fixity of the Past." *Philosophical Review* 121: 179–207.

Jackson, Frank. 1977. "A Causal Theory of Counterfactuals." *Australasian Journal of Philosophy* 55: 3–21.

Khoo, Justin. n.d. "On Indicative and Subjunctive Conditionals." Unpublished manuscript.

Kment, Boris. 2006. "Counterfactuals and Explanation." *Mind* 115: 261–309.

Kutach, Douglas N. 2002. "The Entropy Theory of Counterfactuals." *Philosophy of Science* 69: 82–104.

—. 2007. "The Physical Foundations of Causation." In Price and Corry (2007): 327–50.

—. 2013. *Causation and its Basis in Fundamental Physics*. Oxford: Oxford University Press.

Lange, Marc. 2000. *Natural Laws in Scientific Practice*. Oxford: Oxford University Press.

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

—. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13: 455–76. Reprinted in Lewis 1986: 32–66.

—. 1981a. "Are We Free to Break the Laws?" *Theoria* 47: 113–21. Reprinted in Lewis 1986: 291–8.

—. 1981b. "Causal Decision Theory." *Australasian Journal of Philosophy* 59: 5–30. Reprinted in Lewis 1986: 305–39.

—. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61: 343–77. Reprinted in Lewis 1999: 8–55.

—. 1986. *Philosophical Papers, vol. 2*. Oxford: Oxford University Press.

—. 1994. "Humean Supervenience Debugged." *Mind* 103: 473–90. Reprinted in Lewis 1999: 224–47.

—. 1999. *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.

Loewer, Barry. 2007. "Counterfactuals and the Second Law." In Price and Corry 2007: 293–326.

—. 2012. "Two Accounts of Laws and Time." *Philosophical Studies* 160: 115–37.

Maudlin, Tim. 2007. *The Metaphysics Within Physics*. Oxford: Oxford University Press.

McKay, Thomas and Johnson, David. 1996. "A Reconsideration of an Argument against Compatibilism." *Philosophical Topics* 24: 113–22.

Nolan, Daniel. 1997. "Impossible Worlds: A Modest Approach." *Notre Dame Journal of Formal Logic* 38: 535–72.

Norton, John. 2007. "Causation as Folk Science." In Price and Corry 2007: 11–44.

Nute, Donald. 1980. *Topics in Conditional Logic*. Dordrecht: Springer.

Placek, Tomasz and Müller, Thomas. 2007. "Counterfactuals and Historical Possibility." *Synthese* 154: 173–97.

Poincaré, Henri. 1896. *Calcul des Probabilités (Probability Calculus)*. Paris: Gauthier-Villars.

Pollock, John. 1976. "The 'Possible Worlds' Analysis of Counterfactuals." *Philosophical Studies* 29: 469–76.

Price, Huw. 1996. *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*. Oxford: Oxford University Press.

— and Richard Corry (eds.). 2007. *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited.* Oxford: Oxford University Press.

Reaber, Grant. 2012. "Rational Feedback." *Philosophical Quarterly* 62: 797–819.

Rosen, Gideon. 2002. "The Case for Incompatibilism." *Philosophy and Phenomenological Research* 64: 699–706.

Smith, Peter. 1998. *Explaining Chaos*. Cambridge: Cambridge University Press.

Stalnaker, Robert. 1981 [1972]. Letter to David Lewis. In *Ifs*, ed. William Harper, Robert Stalnaker, and Glenn Pearce, 151–2. Dordrecht: Reidel.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

—. 2000. "Free Will Remains a Mystery." *Philosophical Perspectives* 14: 1-20

Vihvelin, Kadri. 2000. "Libertarian Compatibilism." *Philosophical Perspectives* 14: 139–66.

Wallace, David. Forthcoming. "The Logic of the Past Hypothesis." In *Time's Arrows and the Probability Structure of the World*, ed. Barry Loewer, Eric Winsberg, and Brad Weslake. Cambridge MA: Harvard University Press.

Werner, Thomas. 2003. *Deducing the Future and Distinguishing the Past: Temporal Interpretation in Modal Sentences in English*. Ph.D. Thesis, Rutgers University.

Williamson, Timothy. 2007. *The Philosophy of Philosophy.* Oxford: Oxford University Press.

Wilson, Jessica M. 2014. "Hume's Dictum and the Asymmetry of Counterfactual Dependence." In *Chance and Temporal Asymmetry*, ed. Alastair Wilson, 258–79. Oxford: Oxford University Press.

Notes

[1] Thanks to Frank Arntzenius, Jeremy Goodman, John Hawthorne, Boris Kment, David Manley, Kieran Setiya, and David Wallace; to audiences at Columbia, Cornell, and the University of Massachusetts, Amherst; and to an anonymous referee for this journal.

[2] I take laws to be propositions, and treat 'The proposition that P is a law' and 'It is a law that P' as equivalent. Note that laws need not be propositions *about* lawhood. It is plausible that only true propositions can be laws, but I will not assume this since Lange (2000) and Kment (2006) have recently embraced the possibility of false laws.

[3] It might be useful to distinguish my 'determinism' from three nearby doctrines which also have a strong claim on the name. To clarify their logical relations, note that determinism is equivalent to the following claim:

> Any set of propositions that contains all true laws of nature and at least one true history-proposition entails every proposition with which it is consistent.

This can be strengthened by dropping the second 'true': call the result 'general determinism'. In possible-worlds terms, understanding a "nomically possible" world as one where all the true laws are true, general determinism says that any two nomically possible worlds that agree on some history-proposition agree on every proposition, whereas determinism only requires this in the case where one of the two is the actual world.

The other two doctrines can be derived from determinism and general determinism, respectively, by dropping the first 'true'. Of course this makes no difference unless falsehoods can be laws. Lange (2000) and Kment (2006) think that falsehoods can be laws, and favour weaker definitions of determinism that drop my restriction to true laws. However neither of them argues that there actually *are* any false laws, so their view does not undermine the interest of determinism in my stronger sense.

Further notions of determinism can be derived from those just considered by restricting some or all of the quantifiers to *qualitative* propositions and properties. I will say more about these in §2.

[4] Given *Closure* we can strengthen the conclusion further: since every proposition is metaphysically necessitated by a contradiction, *every* counterfactual of the form 'If x had blinked at t, it would have been the case that P' must be true. For example, if x had blinked at t, x would have turned into a butterfly and flown away.

[5] Note that the following also sounds obviously true:

(2′) If determinism is true and x does not blink at t, then if x had blinked at t, a miracle would have to have occurred.

Although I hold that *Past* fails in ordinary contexts, I am inclined to think that (2′), like (2), is true in the context it most naturally evokes. Lewis's dichotomy between 'backtracking' and 'standard' contexts is not particularly helpful here. I believe the explanation turns on subtle ways in which epistemic necessity modals (like 'have to') can serve to signal that certain other propositions, serving as premises from which the asserted content can be inferred, are to be taken for granted.

[6] Explicit endorsements of *Triviality* have not been common. In the tradition of "branching time" (or "branching spacetime") model theory, some theorists (Werner 2003, Placek and Müller 2007) give semantics for counterfactuals which make them equivalent to the corresponding material conditionals in models where there is no branching of histories. Moreover, in this tradition 'determinism' is sometimes used for a thesis true in all and only non-branching models. However, it is not clear that this pins down a unique meaning for 'determinism'; and even if it did, it would be unclear whether it entails determinism in the sense we are concerned with. For an example of the confusions to which this use of 'determinism' can give rise, see Placek and Müller (2007: 3), who accuse Lewis of 'double talk' for taking determinism to be consistent with the (metaphysical) possibility of miracles.

[7] Hawthorne (2006) gives a closely related argument for the implausibility of determinism (what he calls '*de re* determinism').

[8] Equivalently: Every qualitative proposition is such that either it or its negation follows from every set that (a) contains every true law of nature and (b) contains a true proposition of the form *some initial segment of history has F* for some property F that entails every intrinsic property with which it is consistent. Other versions of 'qualitative determinism' can be extracted from this by: (i) deleting 'true' in front of 'law of nature' in (a) (which is a genuine weakening only given the unpopular view that laws can be false); (ii) adding 'qualitative' in front of 'law of nature' in (a) (which is a genuine strengthening only given the unpopular view that laws can be non-qualitative); (iii) adding 'qualitative' immediately before 'intrinsic' in (b) (which is a genuine strengthening given the popular view that there are non-qualitative intrinsic properties); (iv) deleting the last 'true' (which is uncontroversially a genuine strengthening, analogous to the step from determinism to "general determinism" discussed in note 2). Lewis's preferred definition of determinism (Lewis 1984) in effect makes both strengthening (iii) and strengthening (iv). Hawthorne's 'qualitative determinism' makes strengthening (iii) but not (iv).

[9] For this objection to *General Closure*, see Nolan 1997. For a defence of the view that counterfactuals with impossible antecedents are vacuously true, see Williamson 2007, ch. 4.

[10] This is because of Lewis's rejection of the 'limit assumption', according to which for any world w and set of worlds S, some member of S is at least as close to w as any other member of S.

[11] Pollock (1976) and Fine (2012), among others, have forcefully objected to Lewis, arguing that infinite agglomeration is just as unproblematic as its finite analogue.

[12] Fine (2012) develops a logic for counterfactuals, one version of which allows $P > (Q \vee \neg Q)$ to be false even when P is metaphysically possible. (Here > stands for the counterfactual conditional). If this logic is taken to characterise the intended reading, it will require rejecting *Weak General Closure*. The motivation Fine gives for this logic turns on the suggestion that 'If it were to rain in Peoria, the party would be a roaring success' ($P > R$) has what he calls a 'connectionist' reading, under which it is false if the weather in Peoria has absolutely no bearing on whether the party is a roaring success. Fine maintains that the falsehood of this sentence is consistent with the joint truth, on the same connectionist reading, of 'If it were to rain in Peoria and Quentin came to the party, it would be a roaring success' ($(P \wedge Q) > R$) and 'If it were to rain in Peoria and Quentin didn't come to the party, it would be a roaring success' ($(P \wedge \neg Q) > R$). Since $P > R$ follows from $(P \wedge Q) > R$, $(P \wedge \neg Q) > R$ and $P > (Q \vee \neg Q)$ given other intuitive logical principles accepted by Fine, he concludes that $P > (Q \vee \neg Q)$ is false. However, I see no strong motivation for positing the 'connectionist' reading. The obvious conservative account of the example is that when we know that the party will be a roaring success no matter what happens in Peoria, $P > R$ is still true but pragmatically disfavoured (perhaps because of the desirability of putting 'still' into the consequent). Anyway, neither *Past* nor *Laws* has much pre-theoretical plausibility on a connectionist reading: intuitively, a blink at *t* has no more bearing on the laws or the distant past than the weather in Peoria has on the success of Fine's party. So the CCA only presents us with an interesting paradox on a non-connectionist reading of the counterfactuals, for which Fine has given no reason to reject *Closure*.

[13] Other friends of counterfactual miracles include Jackson (1977), Halpin (1991), Vihvelin (2000), Beebee (2003), Woodward (2003, §3.5), Lange (2000), Kment (2006), Glynn (2013), and Khoo (n.d.). Note that while Lange and Kment agree with Lewis that if determinism is true and x doesn't blink at t, some laws are such that they would not have been *true* if x had blinked at t, they hold that in a different sense the laws do not counterfactually depend on what we do: if a proposition is a law, it would still have been a law if x had blinked at t, although it might in that case have been a false law rather than a true one. This wrinkle does not undermine the arguments for *Laws* that I will be giving in sections 6 and 7.

[14] Other proponents of law-preserving views of counterfactuals under determinism include Nute (1980), Bennett (1984) (although he expresses second thoughts in Bennett 2003), Goggans (1992), Albert (2000, 2015), Loewer (2007), Kutach (2002, 2007), and Wilson (2014). The discussion in the present section is especially indebted to Albert and Loewer.

[15] *Holding History Fixed* is not about how we evaluate counterfactuals whose consequents are *primarily about* times earlier than the times the antecedent is primarily about, like 'If John had forgotten to have breakfast this morning, he would still have had breakfast yesterday morning'. On the rare occasions when we consider such counterfactuals, we often resolve their context-sensitivity in a non-standard way that involves holding fixed less of the truth about prior history (Lewis's 'backtracking' interpretations). The case for the prevalence of *Holding History Fixed* is based on conditionals like (5–7) whose consequents entail certain propositions about earlier times without being primarily about those times. Indeed, given the popular view that the very existence of particular people and ordinary objects metaphysically necessitates quite a lot about the course of earlier history, almost all ordinary counterfactuals will fall into this category. Note that this contradicts Maudlin's suggestion (2007, chapter 1) that given determinism, counterfactuals whose antecedents are about a particular time t can typically be evaluated by a recipe that involves making minimal changes to the instantaneous state at t so as to make the antecedent true, evolving forwards in accordance with the laws, and stopping once we have come up with a description of a long enough interval beginning at t to entail the truth or falsity of the consequent.

[16] The 'exploding difference' view considered by Bennett (2003: 218) fits this template. However, Bennett incautiously glosses 'approximate match' as match 'in every respect we would ever notice or think about'. Taken literally, this is too demanding. As Ahmed (2014: ch. 5) points out, we are capable of thinking about the question whether the state of the universe a million years ago was such as to entail, together with a certain proposition L, that we act in a certain way. In the case where L is in fact a true, deterministic law, proponents of *Laws* must agree that *this* question would have had a different answer if we had acted differently, despite being entirely about the distant past.

[17] Halpin (1991) expresses the same concern.

[18] *Proof*: Suppose for contradiction that this is false: for some $p \in M$ and positive real numbers t and ε, every open neighbourhood O of p contains a point p' such that for some $x \in [0,t]$, $d(\Phi(x,p'),\Phi(x,p)) \geq \varepsilon$. Then, by taking any nested sequence of open sets O_1, O_2,… whose intersection is $\{p\}$, we could construct a sequence $\langle x_1,p_1\rangle$, $\langle x_2,p_2\rangle$, … in $\mathbb{R} \times M$ such that p_1, p_2, … converges to p; each x_i is in $[0,t]$; and $d(\Phi(x_i,p_i),\Phi(x_i,p))$ is always $\geq \varepsilon$. Since $[0,t]$ is a closed interval, the Bolzano-Weierstrass theorem entails that the sequence x_1, x_2, … has a subsequence x_{k_1}, x_{k_2}, … that converges to some limit point $y \in [0,t]$. Then by the continuity of Φ, $\Phi(x_{k_1},p)$, $\Phi(x_{k_1},p)$,… converges to $\Phi(y,p)$. Moreover, since p_{k_1}, p_{k_2}, … converges to p (being a subsequence of p_1,p_2…), the sequence $\Phi(x_{k_1},p_{k_1})$, $\Phi(x_{k_1},p_{k_1})$, … must also converge to $\Phi(y,p)$. Hence by the continuity of d, the sequence $d(\Phi(x_{k_1},p_{k_1})$, $(\Phi(x_{k_1},p))$, $d(\Phi(x_{k_2},p_{k_2}),(\Phi(x_{k_2},p))$, … converges to $d(\Phi(y,p),\Phi(y,p))=0$. This contradicts the assumption that $d(\Phi(x_i,p_i),\Phi(x_i,p)) \geq \varepsilon$ for every $\langle x_i,p_i\rangle$. (Note that this argument does not actually require the similarity-measure d to be a metric: it is enough for it to be a continuous function for which $d(p,p)$ is always zero.)

[19] Smith (1998) provides a philosophically accessible introduction to chaos theory.

[20] In fact almost all of E will consist of states corresponding to an earlier history *radically unlike* that of the actual world—a history much of which looks rather like the post-t history of the actual world played in reverse. See Albert 2000, ch. 4.

[21] To be precise: for any $S \subset M$, let $\Phi(x,S) = \{q: \exists p \in S(q = \Phi(x,p))\}$ be the result of evolving all points in S by x units of time. Let u be some small (but not too microscopically small) positive real number, representing a unit of time. Where Θ^- is a finite set of ordered pairs of negative multiples of u and macrostates, and Θ^+ is a finite set of ordered pairs of positive multiples of u and macrostates, let $\Xi^- = \cap\{\Phi(x,S) \mid \langle x,S\rangle \in \Theta^-\}$, and $\Xi^+ = \cap\{\Phi(x,S) \mid \langle x,S\rangle \in \Theta^+\}$. Finally, let M_0 be some macrostate, and for any measurable $S \subset M$, let $P_{M_0}(S)$ equal the volume of $S \cap M_0$ (according to the natural measure) divided by that of M_0. Then the Independence Conjecture says that $P_{M_0}(\Xi^- \cap \Xi^+) \approx P_{M_0}(\Xi^-)P_{M_0}(\Xi^+)$.

The Independence Conjecture is a special case of the 'Simple Dynamical Conjecture' discussed by Wallace (forthcoming), where the role of the 'coarse graining map' is played by the function that takes each probability distribution on M onto the distribution that agrees with it on each macrostate but is uniform within macrostates, and the 'simple distributions' are the uniform distributions over particular macrostates.

[22] An interesting question is whether the empirical success of statistical mechanics merely supports the claim that the conjecture is true as regards the true underlying microdymamics, whatever it might be (if determinism is true), or whether it also supports the kind of mathematical intuition that would leads us to assume that such behaviour is typical of complex worlds with simple deterministic laws.

[23] The following may help to make this intuitively plausible: there is some finite n such that, by specifying the positions and momenta of all particles up to the first n decimal places, we can guarantee the generation of the puff of wind that types the first character of *Hamlet*. But as time goes on, more and more digits in the initial positions and momenta will become macroscopically relevant So insofar as it is plausible that there is an assignment of values to the first n digits that results in the first letter being typed, it should also be plausible that for some $n'>n$, there is an assignment of values to the first n' digits that result in the first two letters of *Hamlet* being typed…. We need not worry that the fine-tuned choices of parameters required to generate the first thermodynamic miracle will tie our hands in such a way as to make subsequent thermodynamic miracles nomically impossible, since each such miracle only requires fine-tuning finitely many decimal places' worth of the initial state, and we have infinitely many decimal places to play with.

[24] One might worry that if we allow for an earlier divergence in interpreting Fred's utterance, we will be forced to say that the proposition he asserted was quite improbable on his evidence, on the grounds that among the worlds where Fred is with Mary at t and the course of history up to the relevant earlier time approximately matches actuality, most (in the sense of the Lebesgue measure) are not worlds where Fred hugs Mary just after t—rather, they are worlds where the conversation does not take place at all, or takes place in a different way and on a different schedule. But this assumes that in evaluating a counterfactual, we can't hold fixed *any* truths about a time unless we hold fixed the complete approximate truth about that time. And there is no reason to accept this. In interpreting Fred's utterance, we are free to hold fixed both approximate history up to, say, one day before t, and also the facts about what whatever Mary said just before t that inspired Fred's impulse to give her a hug.

[25] The seminal work here is Poincaré's (1896) analysis of roulette.

[26] Moreover, the ratios of the volumes of these sets of points will be approximately proportional to the chances that our high-level chancy theories assign to the outcomes they yield, a fact which helps explain the success of these theories.

[27] We will need to allow for the relevant source of context sensitivity to be bound by some higher quantifier, to account of sentences like 'Any time a moving atom collides with a stationary atom, the collision would not have taken place if the trajectory of the moving atom had different even by a fraction of a degree'. Such "bindability" seems to be a typical feature of context-sensitive expressions: see the discussion of 'local' in Dorr and Hawthorne 2014, §3.3.

[28] Unlike Albert and Loewer, the most prominent defenders of a law-preserving approach to counterfactuals under determinism, I have not tried to explain why statistical-mechanical probabilities should play a role in determining our credences in counterfactuals. I think that a story about this can be developed, but I will not tell it here since I do not think it is necessary for defending *Laws*. My view is also unlike that of Albert and Loewer in that I think we should generally hold fixed the entire approximate course of history up to some contextually given time, whereas they merely hold fixed the macrostate at the relevant time and the "Past Hypothesis" (roughly, the macrostate at the beginning of the universe). This leads them to odd results: for example, that 'If I had raised my arm, Atlantis would never have existed' deserves high credence, conditional on the hypothesis that Atlantis existed but left no macroscopic traces (Kutach 2002; Loewer 2007, 2012; Albert 2015, ch. 1). They are motivated to tolerate these results by the desire to avoid writing any explicit time asymmetries into the semantics of counterfactuals. But given that my more conservative approach is consistent with all manner of reductive ways of distinguishing past and future—e.g. in terms of the entropy gradient—I see no deep reason to follow Albert and Loewer here.

[29] Of course, given the existence of backtracking readings of counterfactuals, no-one should suppose that this connection between causation and counterfactuals obtains on *all* readings. But the objector can insist that it obtains on *some* important readings, including those we have been focusing on.

[30] Given that Lewis is also the most prominent defender of a theory of causation on which counterfactual dependence is sufficient for causation, how can he possibly reconcile his embrace of short-range backwards counterfactual dependence with the denial of backwards causation? His strategy turns on two of his other commitments: (i) the rejection of Conditional Excluded Middle (the schema 'Either if it were that P it would be that Q, or if it were that P it would be that not-Q') and (ii) his sparse account of events, according to which many propositions are not necessarily equivalent to the occurrence of any event:

> That is not to say, however, that the immediate past depends on the present in any very definite way. There may be a variety of ways the transition might go, hence there may be no true counterfactuals that say in any detail how the immediate past would be if the present were different. I hope not, since if there were a definite and detailed dependence, it would be hard for me to say why some of this dependence should not be interpreted—wrongly, of course—as backwards causation…. (Lewis 1979: 40)

If this strategy for blocking backwards causation worked for Lewis, there is no obvious reason why it should not work just as well for proponents of *Laws*. However, I can't see how Lewis's hope could be justified given his own account of counterfactuals, according to which the closest worlds are those with the latest possible small miracle. Suppose a bomb explodes at *t* (an event), and the flimsy building it is in collapses (another event). There are presumably worlds which diverge from actuality by a small miracle just before *t* where neither event occurs. But for a small miracle to lead to a world

[31] The 'big miracle' view seems to be implied in Jackson (1977). It is explicitly defended by Glynn (2013).

[32] For a more elaborate version of this argument, see the discussion of 'bumps' in Bennett 2003, §79.

[33] On one version of the view, the relevant possible worlds would discontinuously splice together the pre-noon segment of the actual world with the post-noon segments of unremarkable, nomically possible worlds where I walk to the car in the usual way.

[34] Note that giving up the sufficiency claim need not mean giving up the project of analysing causation in counterfactual terms. For one thing, such analyses could be framed using interpretations of counterfactuals distinct from the "standard" ones we have been concerned with. For another, most recent attempts to analyse causation in counterfactual terms (e.g. Hitchcock 2011) end up testing for causation using arrays of counterfactuals with quite complex antecedents, which vary the putative cause while also explicitly holding fixed a wide array of other facts about relevant parts of the world. There is no obvious reason to expect analyses of this form to vindicate the sufficiency claim.

[35] For various developments of this theme, see Price (1996), Field (2003), and Norton (2007).

[36] The argument here and in the rest of the present section amplifies the brief presentation in section 3.2 of Dorr and Hawthorne 2014.

[37] On some views of propositions, we will be able to grasp and express these complex propositions under certain modes of presentation. For example, suppose we introduce a new predicate 'actualistic' by stipulating that it should express whatever intrinsic profile is instantiated by the first million-year segment of the history of our universe. On the views in question, 'The initial million-year segment is actualistic' will end up expressing a history-proposition that nomically necessitates all truths (assuming determinism). But this does not lead to counterfactuals of the problematic sort. If history had been different, 'actualistic' would have expressed a different property; so if our protagonist's only way of believing the relevant history-proposition goes by way of this predicate, she would still have had true beliefs even if history had been slightly different.

[38] Arguments of the present sort can, of course, be used to support the counterfactual robustness of *simple* truths about the past. This has some controversial consequences when combined with the counterfactual robustness of the laws. For example, suppose that the initial state of the universe was mirror-symmetric (a simple truth), and that the deterministic laws are parity-invariant so that things remain symmetric forever. If we hold all this fixed, we will have to endorse many counterfactuals to the effect that if we had done such-and-such, our mirror-counterparts would also have done such-and-such. Given the apparent lack of causal relations between mirror-counterparts, these counterfactuals are controversial, especially if we make the further, un-Humean stipulation that the symmetry of the initial state was nomologically contingent. Nevertheless, my inclination is to follow the arguments where they lead, and chalk the case up as another counterexample to the sufficiency of counterfactual dependence for causation.

Here I am disagreeing with Boris Kment and an anonymous referee, who argued from the falsity of the controversial counterfactuals in the symmetric world to the general untrustworthiness of the intuitions underlying our aversion to (8) and (9). Even granting the premise, the conclusion strikes me as an overreaction. I think it would be better methodology to resolve conflicts among our intuitive judgments about more exotic possibilities (like the symmetric) world by extrapolating general principles (such as *Laws*) from our judgments about less exotic possibilities (like the Frank case).

[39] More realistic hypotheses of this kind arise in connection with certain many-worlds interpretations of quantum mechanics.

[40] Moreover, if people at the worlds in question have such rich essences, our counterfactual judgments concerning those worlds will be *riddled* with errors, so that the worlds present no *distinctive* challenge for the mode of reasoning that leads us to judge (8) and (9) false in the original example.

[41] Goodman (2015) gives an epistemological argument for the counterfactual fragility of the past that nicely complements the broadly ethical arguments of the present section. His premises are, first, that some deterministic propositions can be known, and second, that one could not know any proposition that was counterfactually fragile in the way that any deterministic proposition would have to be given *Past*. The concept of lawhood plays no role in this argument, although the best candidates to be knowable deterministic propositions are laws.

[42] If I am right about this, then probably, if Bill has a realistic sense of the extent of counterfactual dependence across his life, hardly *any* of the mistakes he made before meeting Mary will be things he can appropriately regret. More generally, people whose lives have gone well should have few regrets about their early choices. Of course we often do have such regrets, and judge them appropriate. Nevertheless, I do not think that the claim that they are mostly inappropriate amounts to a radical error theory about our ordinary normative beliefs. The disposition to let go of regret when we reflect on the relations of counterfactual dependence between our mistakes and the things we most value is deeply ingrained in our ordinary practice, even though we do not always remember to engage in such reflection. (Thanks to an anonymous referee for pressing me on this point.)

[43] The connection I have suggested between regret and counterfactual thinking is probably too simple. For example, can't a reformed criminal wholeheartedly regret having committed a certain crime even when he knows that because of his past depravity, if he had not committed that crime, he would have chosen to commit an even worse crime instead? This suggests a revision along the following lines: one cannot rationally regret doing something (all things considered) if one rationally believes that for each alternative to doing that thing, things would have been worse if one had performed that alternative. But note that even this weaker principle will tell us that Frank's regrets are irrational.

[44] Since causal decision theory merely requires a certain kind of fit between one's choice, one's values, and one's credences in counterfactuals, it is consistent with different views about how one should form credences in the relevant counterfactuals. It is not uncontroversial that it would be rational to form them just by induction from one's past track record: for example, Arntzenius (2008) claims that in situations where multiple predictions about one's future behaviour would be self-fulfilling, one rationally ought to form credences optimistically, predicting that one will do whatever one most *hopes* to do, in spite of putative countervailing evidence such as a bad track record.

[45] For the argument to go through, she does not actually need to be *certain* that there is a history-proposition which would be true no matter what she did. It is enough if her credence that this is not the case is less than her credence that L would be false no matter what she did.

[46] Lewis's version of causal decision theory (Lewis 1981b) would also serve the purposes of the argument, since it coincides with the Stalnaker-Gibbard-Harper theory in deterministic worlds.

[47] Ahmed gives several other arguments for evidential decision theory, but here I will confine my attention to the argument from *Betting on the Laws*.

[48] Ahmed also thinks that only theories that capture this intuitive thought deserve the name 'causal decision theory'. I have been using this label in a way that Ahmed would oppose, but nothing turns on this terminological issue.

[49] These are the adjectives Ahmed (2014: 140) uses in discussing the combination of the Stalnaker-Gibbard-Harper decision theory with Loewer's (2007) statistical-mechanical approach to counterfactuals. I assume the accusation is intended to extend to other views that preserve *Laws*.

[50] My main worry about (i) is that if one interprets the slogan about the relevance of causation very literally, *no* decision theory that applies nontrivially in deterministic worlds can endorse it. As we saw in §5, Lewis denies that the miracles which would have occurred if we had acted otherwise would have been *caused* by (or constituted by, or identical to) our actions. So if Lewis is right, *both* the initial conditions *and* the laws would seem to be "beyond our present influence", causally speaking; but of course no decision theory that required treating both the initial conditions and the laws as "states of nature" could apply nontrivially under determinism. This suggests that we need to be more flexible in interpreting the slogan. As Ahmed actually interprets it, it merely requires treating *particular matters of fact* causally independent of the agent's choice as "states of nature", where universal generalisations (including laws) do not count as particular matters of fact. While this is certainly a possible interpretation, it is not obviously the only one, or the best one.

[51] In defending his characterisation of the Stalnaker-Gibbard-Harper-Loewer view is 'unstable', Ahmed at one point (2014: 141 n. 26) seems to take it for granted that the "everyday counterfactual" is one for which the past is to be held fixed, so that the view he is arguing against is one according which the conditionals relevant to deliberation differ in this respect from everyday counterfactuals. This combination of views is certainly problematic: given how important decision-making is in our lives, it would be mysterious if natural languages lacked the resources to express the kind of conditional that matters to rational decision-making.

[52] Ahmed also has another case, *Betting on the Past*, which can play a similar dialectical role.

[53] Instead of 'x cannot make it the case that p is false', Ginet says 'it is not open to x to make it the case that p is false', where 'open to' has a rather intricate further definition.

[54] Here I believe I am in agreement with van Inwagen (2000). In that paper, van Inwagen, in effect, rejects version (i) while endorsing version (vi) as capturing what version (i) was 'trying to capture'. But there is evidence that he takes (vi) to be equivalent to (ix), namely his taking the thesis that 'no non-actual world is as close to the actual world itself' to entail that he is able to ensure the truth of a proposition that it is true only at the actual world (note 9).

[55] Here is a counterexample to *Closure* in version (vii) of the SCA. Suppose that at t Polly can raise her arm; that Quentin is asleep at t'; and that Polly at t has no control in any sense over whether Quentin is asleep at t'. Then each of the things Polly can do at t is such that if she had done it, her doing it would neither be nor cause any event inconsistent either with the proposition that Quentin is asleep at t', or with the proposition that (Polly raises her arm if and only if Quentin is awake at t'). Nevertheless, something Polly can do at t is such that if she had done it, her doing it would have been inconsistent with the proposition that Polly does not raise her arm.

[56] Opponents of Conditional Excluded Middle should regard the following (based on an example in McKay and Johnson 1992) as a counterexample to version (viii) of *Closure*. Let p be the proposition that a certain fair coin never lands Heads, and q the proposition that it never lands Tails. The coin is never tossed, so both p and q are true. At t, Cecil was able to toss the coin. Even though the coin would have landed Heads or Tails if he had tossed it, it is neither the case that it would have landed Heads if he had tossed it nor that it would have landed Tails if he had tossed it. (This stipulation requires a failure of Conditional Excluded Middle.) More generally, nothing Cecil was able to do at t was such that if he had done it, p would have been false, or such that if he had done it, q would have been false. Nevertheless, one thing that Cecil was able to do at t—namely, tossing the coin—is such that if he had done it, the conjunction of p and q would have been false. Proponents of Conditional Excluded Middle will reject this counterexample, but for them version (viii) of the SCA is equivalent to version (ix) (given the uncontroversial claim that no normal person can do anything such that a contradiction would be true if they did it).

[57] Suppose for reductio that determinism is true, and at t, x is normal and able to ϕ but does not in fact ϕ. By *Triviality*, the proposition that x does not ϕ would still have been true if x had ϕed. So a contradiction would have been true if x had Φed. So by the uncontroversial claim, x is not able to ϕ: contradiction.

[58] It has sometimes been suggested that the only viable motivation for *Laws* in the ACA and related arguments depends on an anti-Humean metaphysics of lawhood, so that Humeans have no reason to accept this premise (see Beebee and Mele 2002 and Rosen 2002). If this were true, it would also be true for *Laws* in the CCA. But the arguments of sections 6 and 7 show that there are arguments for *Laws* that have nothing to do with anti-Humeanism. Indeed, as far as I can tell, my judgment that Frank's career would not have been devoted to a mistake if we had put a glass of water on the lectern was driven not so much by the stipulation that L was a law, as by the very features of L that constitute lawhood according to Humeans like Lewis (1994), namely its qualitativeness, its simplicity, and its strength.

[59] Van Inwagen (1983) and Ginet (1990) both rely heavily on abductive arguments based on particular examples in which people seem to have no choice about the truth of particular propositions about the laws and the past.

[60] Although it is certainly not unprecedented. For example Vihvelin (2000), although a friend of counterfactual miracles, broadly agrees with me about the dialectical relevance of the CCA to the ACA.

[61] For example, Beebee (2003: 260) takes it for granted that if determinism is true and I do not raise my hand, then 'raising my hand would require a miracle', while Holliday (2012) takes it for granted that 'if s is determined not to do y at t', then if s were to do y at t', the past … would (have to) be different'.