Philosophy of Science (2023), **90**, 77–91
doi:10.1017/psa.2022.14

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

Backward Causation: Harder Than It Looks

Athamos Stradis

King's College London, London, UK
Email: athamos.stradis@kcl.ac.uk

(Received 11 August 2020; revised 27 November 2021; accepted 03 March 2022; first published online 22 April 2022)

Abstract

According to David Albert, there are certain situations where we can cause events that lie in our past. In response to a well-known objection that we never observe backward causation, he argues that there are good reasons why we can't tell when it obtains. However, I identify another difficulty with Albert's view: at face value, it has the unattractive consequence that backward causation is not just possible, but rife. In this article, I show how this implication can be blocked. I then use my analysis to defend Albert's account from a second well-known objection, namely, that it allows us to control the past.

1. Introduction

Albert (2000, 2014) has offered a broadly Lewisian account of causation in which the causal arrow ultimately stems from the record asymmetry—the fact that we have records of the past and not the future. A surprising corollary is that there are certain scenarios where we can cause events that lie in our past. At first sight, this might seem inconsistent with the fact that we never observe backward causation. But by showing that this phenomenon is by its nature unobservable, Albert takes this criticism in stride.

Nevertheless, I identify another difficulty. We ordinarily think backward causation is rare or nonexistent. Although Albert suggests his account respects this widely held view, at face value it does not, for the following reason. In his account, any unrecorded past event is vulnerable to backward causation. But one would think that there are countless unrecorded minor events in our past, from coconuts falling in long-lost jungles to pterodactyls squawking in Jurassic skies. If this is right, then countless events are vulnerable to backward causation. This implies that backward causation is not just possible, but rampant throughout the world. Lewisian accounts have long been suspected of having this feature, and this paper highlights new grounds for suspicion in the case of Albert's account in particular.[1]

[1] Kutach (2002) provides other arguments to the effect that Lewisian accounts based on statistical mechanics, Albert's being an example, allow backward causation.

To remedy this, I propose a mechanism whereby records persist while *seeming* to vanish. The implication is that despite appearances, there are *not* countless unrecorded minor events in our past, so backward causation is genuinely rare/nonexistent after all. Not only does this square Albert's causal account with common sense, it also disarms a second well-known objection relating to control.

This article is structured as follows: In section 2 I outline Albert's truth conditions for causal statements, and explain how we have probabilistic epistemic access to these via the "Mentaculus." In section 3 I explain how this picture is meant to yield a causal arrow that usually points forward, but occasionally backward. I also show how Albert responds to a well-known objection that we never observe backward causation: he claims it is unverifiable. In section 4 I identify a different problem: when Albert's account is paired with commonsense ideas about records, it has the unattractive consequence that backward causation is extremely common. In section 5 I revise these commonsense ideas to overcome this issue. In section 6 I use my response to disarm a second well-known objection to Albert, that says his account allows us to control the past. I conclude in section 7.

2. Causation in the Mentaculus

Albert wants to show why we are "likely to be in a position to influence much about the future and next to nothing about the past" (2014, 165). In context, it is clear that Albert uses "influence" to mean "cause." To understand his explanation, we must first come to grips with his framework for causal analysis. This runs as follows: causal statements are cashed out in terms of a factual statement plus a counterfactual statement, and an entity called the "Mentaculus" gives probabilistic verdicts on these.[2] In what follows I lay this all out, starting with causal statements. This isn't exactly how Albert presents his account, but I believe it is a fair reconstruction.

Suppose in our world, some particular event x occurs (I'll always assume this), and we wish to know whether or not it causes some other particular event y in our future/past. In Albert's account, as in Lewis (1973a, 1973b), x causes y if and only if (iff) both of the following causal criteria are satisfied:

> **Criterion 1:** y occurs in our world.
> **Criterion 2:** Had $\neg x$ obtained, $\neg y$ would have obtained.

Criterion 1 is a kind of ground-zero requirement for causation because it establishes that the relevant events actually occur in the first place. Meanwhile, Criterion 2 is the definitive feature of a counterfactual analysis of causation. As Lewis himself recognized, this analysis faces various difficulties and requires some refinement in order to get around them. Two particularly salient problems are preemption scenarios and backtracking counterfactuals.[3] But because these issues aren't directly relevant for us and would only complicate things, I shall set them aside.

[2] Loewer (2007, 2011, 2012) paints a similar picture but analyses causation through the lens of possible human interventions; see Frisch (2007, 353–58) for a comparison. For consistency, I shall only focus on Albert's proposal. See also Kutach (2002) for early formulations along these lines.

[3] See Paul and Hall (2013, chap. 3) and Reiss (2015, 110–14), respectively, for discussions of these issues. See also Kim (1973), who argues that the causal criteria are insufficient to characterize causal connections.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

In summary, a causal statement is determined by a factual statement (Criterion 1) and a counterfactual statement (Criterion 2). So, what dictates their truth values? Within the context of classical statistical mechanics, these are determined by the trajectories picked out by the relevant microstates in the universe's phase space Γ. This is straightforward in the case of Criterion 1: it is satisfied when the *actual* current microstate m_0 time-evolves forwards/backwards under the dynamics to satisfy y. But what is the "relevant microstate" in the case of Criterion 2? Similarly to Lewis (1973b), Albert (2014, 163) consults our closest possible world, which he interprets as the nearest nomologically possible microstate $m_\star$ to m_0 that satisfies $\neg x$ (as measured by distance in Γ). Criterion 2 is then satisfied when $m_\star$ time-evolves forwards/backwards to satisfy $\neg y$.[4] This sums up the truth conditions for causal statements.

Although these truth conditions are conceptually clear, they are epistemically opaque. We can't literally envisage microstates, much less microstates of the *entire universe*, and we certainly can't track their trajectories in Γ. Instead, we must make do with probabilistic verdicts on the causal criteria given by the "Mentaculus." This entity is a combination of three ingredients:

1. The universe's fundamental dynamical laws. For simplicity, these are taken to be Newton's deterministic laws.
2. The "Past Hypothesis," the posit that the universe began in an extremely low-entropy macrostate M_{PH}. Because Albert (2000, chap. 6) regards this as a law, all nomologically possible microstates are consistent with this posit.
3. The "Statistical Postulate," a probability distribution that is uniform (according to the Lebesgue measure) over M_{PH}'s microstates.[5]

The result is a uniform distribution over trajectories in Γ, which Albert regards as the universe's nomologically possible trajectories. By conditionalising the Mentaculus on the actual macrostate M_0 or on the counterfactual macrostate $M_\star$ (where $m_0 \in M_0$ and $m_\star \in M_\star$), we obtain y's respective probabilities in the actual or counterfactual world, *given* all macroscopic evidence in that world.[6] The procedure for finding these probabilities is the same in both cases:

i. Apply a uniform distribution over M_0 or $M_\star$ (this satisfies the Statistical Postulate).
ii. Discard any microstates that didn't evolve from M_{PH} (this satisfies the Past Hypothesis). This leaves us with the nomologically possible sets of present microstates $N_0 \subset M_0$ or $N_\star \subset M_\star$. These are just macrostates conditionalized on M_{PH}.
iii. Time-evolve N_0 or $N_\star$ forwards/backwards under the dynamics to the time of the consequent.

[4] Whereas Lewis (1973a, 1979) envisaged possible worlds satisfying antecedents via "miracles" (violations of law) and identified the closest with recourse to his infamous four-point distance gauge, Albert does away with all this.

[5] Albert (2000, chap. 7) argues that if the Ghirardi–Rimini–Weber (GRW) theory is correct, we can dispense with the Statistical Postulate because the universe's chanciness will be built directly into the dynamical laws. But because we're assuming a classical framework, we can put this idea aside.

[6] Of course, we never observe *all* this evidence. But presumably, observing a large chunk of it brings our credences into rough agreement with the Mentaculus' probabilities.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

iv. Determine the proportion of N_0 or $N_\star$ that find themselves in a macrostate in which y occurs. There may be multiple such macrostates, but I shall gloss these as N_y.
v. This gives $\Pr(y \mid N_0)$ or $\Pr(y \mid N_\star)$, that is, the chances of y occurring in the future/past of N_0 or $N_\star$.

This provides us with probabilistic handles on the causal criteria as follows: $\Pr(y \mid N_0)$ is the chance of Criterion 1 being *satisfied*, whereas $\Pr(y \mid N_\star)$ is the chance of Criterion 2 being *violated*. When $\Pr(y \mid N_0) \approx 1$ and $\Pr(y \mid N_\star) \approx 0$, both causal criteria are probably satisfied, and x probably causes y.

3. The Possibility of Backward Causation

We've now identified the conditions under which causation occurs and our means of epistemic access to these conditions via macroscopic evidence. With this groundwork in place, we're ready to see how it yields a causal arrow and why this may occasionally point backwards.

In order to derive the causal arrow, we must first acknowledge that the world exhibits a record asymmetry. At this point, two questions naturally arise. First, what *is* a record? This is notoriously hard to answer in a precise way. But for our purposes, we can stick with what seems to be Albert's own view: records are localied macrostates that are highly informative about other times,[7] and it just so happens that these "other times" always lie in our past.

Second, what explains the record asymmetry? According to Albert (2000, chap. 6), it is the fact that the Mentaculus contains a Past Hypothesis but no analogous "Future Hypothesis." For our purposes, the details and difficulties of this explanation aren't relevant,[8] and I am simply going to take it as read. The implication is that not only the universe's actual macrostate N_0 but also the counterfactual macrostate $N_\star$ contain records of the past and not the future. Through piecemeal observations of *our* world's records—which, recall, are localized macrostates—we are able to partially reconstruct N_0 and $N_\star$, allowing our credences to roughly track the Mentaculus probabilities.

If we grant the foregoing, the causal arrow drops out as follows: As we saw, whether or not x causes y depends on the two causal criteria. If we take Criterion 1 for granted (i.e., both x and y occur in our world), then the question hinges on Criterion 2. This is a counterfactual statement that we gauge via the Mentaculus. Because of the record asymmetry, the Mentaculus gives different answers depending on whether x precedes y, or vice versa, as shown in the following paragraphs.

[7] See Albert (2000, chap. 6) for details. Why are records macrostates and not microstates? I think there is a principled reason for this: in order for something to function as a useful record, it must (i) be reasonably stable over time and (ii) enter into reliable correlations. Microstates, being full specifications of the positions and momenta of all particles, are constantly and erratically changing and therefore lack these key features. Interestingly, Albert's conception of records hinges on what qualifies a macrostate in the first place—a matter than depends on our own physical constitution. See Hemmo and Shenker (2016) for details.

[8] See Frisch (2007) for a useful overview and critique.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

When x precedes y, Criterion 2 is a forward counterfactual. To evaluate this, we time-evolve $N_\star$ forwards and gauge the proportion that lands up in N_y, yielding $\Pr(y \mid N_\star)$. Because N_0 lacks records of the future, it lacks records of y. But because $N_\star$ mimics N_0 except in satisfying $\neg x$, $N_\star$ generally also lacks records of y. It can therefore easily happen that $\Pr(y \mid N_\star) \approx 0$. In such circumstances, Criterion 2 is probably satisfied, and so x probably causes y.

When y precedes x, Criterion 2 is a backward counterfactual. To evaluate this, we time-evolve $N_\star$ backwards and gauge the proportion that lands up in N_y, yielding $\Pr(y \mid N_\star)$. Because N_0 contains records of the past, it generally *does* contain records of y. But because $N_\star$ mimics N_0 except in satisfying $\neg x$, it generally *also* contains records of y. This implies $\Pr(y \mid N_\star) \approx 1$. In such circumstances, Criterion 2 is probably violated, and so x probably *doesn't* cause y.

In summary, here's how the causal arrow emerges in Albert's account. When Criterion 1 is satisfied, x causes y iff Criterion 2 is satisfied. When x precedes y, Criterion 2 is a forward counterfactual; this may well be probable, so x can easily cause y. But when y precedes x, Criterion 2 is a backward counterfactual; this is generally improbable as a result of records, so x probably doesn't cause y.

As Albert (2014, 165–66) observes, the causal arrow that drops out of this isn't absolute; there are certain scenarios where it probably points backwards. To illustrate this, he offers the following example, first discussed by Douglas Kutach.

Suppose Atlantis existed in our past (y) and that I just kept my fingers still (x). Does x cause y? Because Criterion 1 is satisfied, we must turn to Criterion 2. Because y obviously precedes x, Criterion 2 is a backward counterfactual, so we determine it as follows: time-evolve $N_\star$ backwards to antiquity, and gauge the proportion that lands up in an Atlantis-containing macrostate N_y. By hypothesis, N_0 happens to lack records of Atlantis, and because $N_\star$ mimics N_0 except that I just clicked my fingers ($\neg x$), $N_\star$ also lacks records of Atlantis. This implies $\Pr(y \mid N_\star) \approx 0$, which means Criterion 2 is probably satisfied. Therefore, my finger-resting probably causes Atlantis existence. If we generalize from this example, there are certain scenarios where it's probable that we cause events in the past.

Incidentally, this example could be generalized a step further. Although the antecedent here happens to involve human action, this doesn't actually do any important work in the argument. My finger-resting is just a generic present event that could be interchanged for virtually anything, from a leaf falling off a tree to a soap bubble bursting. Therefore, the real upshot is that there are certain scenarios where it's probable that present events *in general* cause past events; *our* doing so is merely a special case of this. But because agency will become important in section 6, I shall stick to examples whose antecedents are human actions.

Kutach's example poses a clear challenge to Albert's account: there are certain situations where we can probably cause past events, and yet we never observe backward causation. How can these ideas be reconciled?

To this end, Albert (2014, 166) draws a key distinction between (a) the chance that backward causation is in play and (b) the chance *we assign* to it being in play, and he rightly points out that these may diverge. This is apparent in the Atlantis case, as follows: On the one hand, we saw that the value of item (a) is 1. This is because it was *stipulated* that Atlantis existed in our past (i.e., that Criterion 1 is satisfied). However, we don't have direct epistemic access to this fact. Instead, our credence

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

comes from conditionalising the Mentaculus on the universe's current macrostate, and because this lacks putative records of Atlantis, it will *seem* like $\Pr(y \mid N_0) \approx 0$. This makes it seem like Criterion 1 *isn't* satisfied, making the value of item (b) negligible. So, even when backward causation is highly probable, we're none the wiser—thus eliminating any verifiable counterexamples to the orthodox view that backward causation is rare/nonexistent.[9]

4. The rampancy of backward causation

Albert's response is good as far as it goes. However, it runs into a different problem. Regardless of whether or not we observe backward causation, we take this phenomenon to be extremely rare or nonexistent. Albert doesn't explicitly flag this as an important feature of causation that he needs to respect. Nevertheless, he seems to acknowledge this implicitly, stressing at various points that we are "likely to be in a position to influence much about the future and next to nothing about the past" (2014, 165). In order to be palatable, Albert's causal account must satisfy the idea that backward causation is rare/nonexistent. At face value, however, it violates this idea, as I explain in the following paragraphs.

In arguing that we can unverifiably cause Atlantis's existence, Albert makes a key hidden assumption: most nomologically possible microstates lacking records of Atlantis didn't contain that city in their past. Without this assumption, we'd have no reason to expect only a tiny proportion of $N_\star$ to time-evolve backwards into N_y, and hence we'd have no reason to expect $\Pr(y \mid N_0)$ to be minuscule, which was the rationale for saying Criterion 2 is probably satisfied in that scenario. Moreover, we'd have no reason to expect only a tiny proportion of N_0 to time-evolve backwards into N_y, and hence we'd have no reason to expect our credence in Atlantis existence to be low, which was the rationale for saying backward causation is unverifiable.

Presumably, this is just a special case of a much more general assumption, which we can express as follows:

Fidelity Assumption: If a nomologically possible region N_i lacks records of y, then $\Pr(y \mid N_i) \approx 0$.

This tells us that if some event occurred but has no records in N_0 or $N_\star$, then the Mentaculus ascribes that event a low probability in both cases. It would then follow a fortiori that if Atlantis existed but has no records in N_0 or $N_\star$, then the Mentaculus ascribes that city's existence a low probability in both cases.

If the Fidelity Assumption is indeed integral to Albert's reasoning, then Atlantis vulnerability to backward causation is just one instance of a wider fact: *all* unrecorded past events are vulnerable to backward causation. As I will explain, this implies something radical about the frequency of this phenomenon.

If the chance of any given event leaving behind no records is miniscule, then there would rarely/never be unrecorded events in our past, so backward causation would be rare/nonexistent, in accord with common sense. There is certainly a large class of *major* events for which this is clearly true, such as the existence of a city, the eruption

[9] This proposed barrier to backward causation is similar to Dummett's (1978). As I understand it, his "dancing chief" example is meant to show that acting for the sake of the past is only worthwhile if we can't know what happened—which we generally *can*, at least in principle.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

of a volcano, or the impact of a large asteroid, all of which we'd expect to leave behind many records. However, it's not obviously true of *minor* events like a stone rolling off a cliff, a Neanderthal's cry, or a carp splashing in a lake. On the contrary, one would ordinarily think such events have a *high* chance of leaving behind no records—at least after an appreciable period of time. Any such event would be subject to backward causation by later events. But because the past is chock-a-block with minor events, this suggests that backward causation is not just possible, but rife.

For clarity, let's see how my objection takes shape with an example. Suppose sometime during the Tang dynasty, a particular carp made a particular splash at a particular time in the Yangtze River (y), and suppose that a moment ago, I kept my fingers still (x). Does x cause y? Because Criterion 1 is satisfied, we must examine Criterion 2 using the Mentaculus: time-evolve $N_\star$ backwards by hundreds of years and find the proportion that lands up in the splashing-carp macrostate N_y. Because N_0 lacks records of the carp's splash, and because $N_\star$ mimics N_0 except that I just clicked my fingers ($\neg x$), $N_\star$ also lacks records of the splash. This implies $\Pr(y \mid N_\star) \approx 0$, which means Criterion 2 is probably satisfied. Therefore, my finger-resting probably causes the carp to splash.

Just as in the Atlantis case, the Fidelity Assumption entails that our credence in the splash is low, so it won't *seem* like my finger-resting causes it. But with a multitude of minor events in our past, it seems we can be confident that there is a vast array of anonymous past events caused by everything we do in the present.[10]

Rampant backward causation follows as a corollary of Albert's causal account plus a commonsense idea about how records vanish, and it contradicts the standard view that backward causation is rare/nonexistent. To deal with this, we face a trilemma: either give up the view that backward causation is rare/nonexistent, revise our ideas about vanishing records, or abandon Albert's causal account altogether. In the next section, I motivate option two.

5. The Dispersion Mechanism

In this section, I propose a mechanism whereby an event's records persist even though they appear to vanish. I shall start by clarifying the statistical relationship between events and their immediate records. Then, by applying these ideas iteratively for later times, I will develop a picture of how events relate to their more temporally remote records.

Any event C raises the chance of a potential record of it, such as A:

$$\Pr(A \mid C) > \Pr(A \mid \neg C). \tag{1}$$

For example, a splashing carp raises the chance of the formation of a circular wave. However, we know from experience that C doesn't guarantee A:

$$\Pr(A \mid C) < 1. \tag{2}$$

In our example, the wave could be suppressed by a sheet of glass or a well-orchestrated gust of wind. What, then, *does* guarantee A's occurrence, prior to that event?

In a deterministic system, every event has a "determinant" at every other moment of time, a "minimal set of conditions jointly sufficient, given the laws of nature, for the fact in question" (Lewis 1979, 474). A determinant is a time-slice of the event's

[10] Like the Atlantis scenario, this is in fact a special case of a more general issue: countless minor past events are caused by any present event we choose to consider, whether it involves human action or not.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

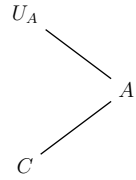

Figure 1. Record *A* drawn with its prior determinant, "*C* plus U_A" (the time axis runs from left to right). Solid lines represent causal correlations.

past or future light cone.[11] *A*'s prior determinant is therefore *C* plus some very complex, far-reaching, microscopically specified set of "background conditions" U_A:

$$\Pr(A \mid C \wedge U_A) = 1. \tag{3}$$

This is illustrated in Figure 1.

From experience, it seems plausible that *C* generally leaves behind multiple records, even though no particular item is guaranteed. This implies that the U_i conditions are not mutually exclusive, and they may be numerous. We can incorporate this into our example by imagining that in addition to U_A leading to the circular wave *A*, there also exists U_B, which leads to a curious heron *B*. This is illustrated in figure 2.

So, given *C*, multiple records tend to form, depending on which U_i conditions happen to obtain. But if we make the assumption that U_i conditions are uncorrelated, the records that end up forming will *also* be uncorrelated—that is, uncorrelated given *C*.[12] The upshot is that in the wake of *C*, the records that *actually* materialize will seem to us to be randomly selected from the set that potentially *could* have materialized, which is to say the set normally associated with *C*.

To understand why this is significant, let's reiterate this logic to see how events like *C* relate to their more temporally remote records. By construing one of these records (say, *A*) as a recordable event in its own right, we can give it the same treatment as *C* itself. This again produces various new records, depending on which U_i conditions obtain. For example, *D* (a startled dragonfly) might result if U_D obtains, whereas *E* (a wobbling reed) might result if U_E obtains. This is shown in figure 3. Whereas A is a "first-generation" record of *C*, we can think of *D* and *E* as "second-generation" records of *C*.

It's easy to see the pattern emerging: *D*, for instance, might cause *F* (well-fed frog) if U_F obtains, and it might cause *G* (awestruck fisherman) if U_G obtains. This is shown in figure 4. Whereas *F* and *G* are immediate or "first-generation" records of *D*, they are "third-generation" records of *C*.

[11] See Arntzenius (1990, 83–84) and Frisch (2005, chap. 8). Lewis (1979) envisaged records playing the role of determinants, but because records are localized, this is inconsistent with the light cone time-slice picture.

[12] What we are looking at here is the famous "fork asymmetry." My assumption that U_i conditions are uncorrelated seems plausible: many have derived the fork asymmetry (or similar structures) from independence conditions in the universe's initial state, and there are various reasons to expect such conditions to hold. For more on these two points, see Frisch (2014) and Lloyd (1994), respectively.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

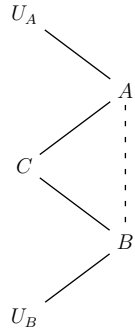

Figure 2. Records *A* and *B* drawn with their prior determinants, "*C* plus U_A" and "*C* plus U_B," respectively. The dotted line represents a noncausal correlation.

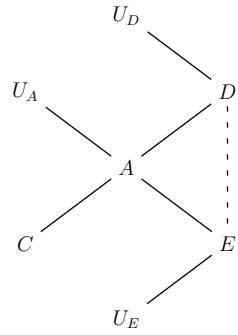

Figure 3. Second-generation records *D* and *E*, together with all background conditions that they depend on.

This "dispersion mechanism" suggests that later-generation records become harder to utilize than early-generation records for two reasons. First, they are more "motley" and hence harder to recognize as records of the original event. Second, even if these records could be recognized as such, they're less reliable in their own right. Let me unpack these claims in turn.

What I call the "motleyness" of later-generation records is a by-product of two things that happen over time. On the one hand, the range of possible later-generation records of *C* becomes larger. This can be appreciated by filling in my omission in figure 4 (see caption). On the other hand, the chance of any given later-generation record becomes smaller because its formation hinges on a longer list of U_i conditions obtaining. For example, whereas *A* hinges on U_A alone, *F* hinges on U_A, U_D, and U_F. Taken together, these two factors ("more diverse" and "less certain to appear") entail that later-generation records have a miscellaneous, motley character, making them less obviously symptomatic of the original event. This is clear in our example: we all recognize a circular wave as indicating a splashing fish, but who can say the same about a well-fed frog?

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

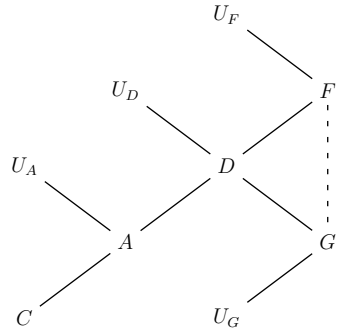

Figure 4. Third-generation records *F* and *G* of *C*, along with all background conditions that they depend on. For simplicity, I've omitted the second-generation record *E* (which branches off from *A*) and the third-generation records it yields.

Meanwhile, the unreliability of later-generation records can be understood as follows: In order to be reliable, a putative record of *C* must have a veridical genealogy stretching back to that event. However, later-generation records have longer genealogies. This is clear from Figure 4: *A* only depends on the genuine occurrence of *C*, whereas *F* depends on the genuine occurrences of *D*, *A*, and *C*. Therefore, when we trace back their origins, later-generation records have more opportunities to be spurious than early-generation records. Our example helps illustrate the contrast: whereas a circular wave misleads us about *C* if it's caused by a falling pine cone (and not the splashing carp), a well-fed frog misleads us about *C* if the frog ate a mosquito (and not a dragonfly), *or* if the dragonfly was startled by a swan (and not the wave), *or* if the wave was caused by a pine cone (and not the carp).

So, here is a short summary of the foregoing. Events generally produce multiple records, none of which is guaranteed, and because U_i conditions are uncorrelated, those records that actually materialize seem randomly realized from the set of possibilities. By reiterating this logic for each of these records, we obtain a picture in which *C*'s late-generation records are motley and unreliable. This means that even long after an event, records still remain, even though they seem to have vanished.

For clarity, let's see how this blocks backward causation. Suppose there's an event *y* in our past. What's the probability of a present event *x* causing it? This is high when the causal criteria are satisfied, and because Criterion 1 is satisfied, we must turn to Criterion 2. However, *y* occupies the role of *C* in this picture, and it therefore leaves records in the present via the dispersion mechanism. Because $N_\star$ also contains these records, $\Pr(y \mid N_\star) \approx 1$, and so Criterion 2 is probably false. This means *x* probably *doesn't* cause *y*. Because this logic applies irrespective of whether *y* is a major or minor event, there *isn't* a vast backlog of past events subject to backward causation. Hence, this phenomenon is genuinely rare/nonexistent after all.

One might raise the following objection. According to this picture, a given very late-generation record of *y* will be extremely unreliable. Therefore, when we conditionalize the Mentaculus on this record, this will raise the probability of *y* only marginally. But in order for it to be unlikely that a present event *x* can cause *y*, we require

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

$\Pr(y \mid N_\star) \approx 1$; that is, we require y's counterfactual probability to be high. How can this be achieved?

The answer is that although records become less reliable with time, this is counterbalanced by the fact that they become more numerous. This follows from the fact that events leave behind multiple records; because this holds true for records themselves, we would expect the sum total of C's records to mushroom with each iterative generation. Again, this is apparent in Figure 4. So, granted that conditionalising on any single very late-generation record raises y's probability only marginally, conditionalising on all of them raises this value significantly. Because $N_\star$ embeds the whole lot, $\Pr(y \mid N_\star) \approx 1$ still stands. So, even long after y has occurred, the chance of a present event x causing it will remain low.

This section comes with a caveat. The dispersion mechanism rests on the notion that events do in fact leave behind multiple first-generation records. However, I did not prove this; I merely stipulated it as plausible. A sceptic might therefore push back by arguing that this isn't true of certain microscopic events—say, a particular Martian dust particle falling in the distant past— which are therefore vulnerable to backward causation. If this doubt is well founded, what would it bode for my account?

My response is that microscopic events like these are not the sorts of ordinary events that usually appear in counterfactuals. Indeed, it looks like the degree to which an event might feasibly fail to leave behind records—and hence fail to be protected from backward causation—is the degree to which it doesn't feature in our standard causal discourse. So, even in this pessimistic scenario, my proposal would still block backward causation of ordinary minor events like splashing carps—and hence still benefit Albert's account, which at face value only avoids backward causation of major events like sunken cities.

6. Implications for control

What does it mean for us to *control* something? It seems clear that in order to control an event, we need to cause it in the first place. At the same time, there are many things that we cause but do not control: happy accidents like roulette wins, unhappy accidents like wine spillages, and the innumerable events that we bring about through our daily activities but aren't even aware of. This implies that whatever control amounts to, it must be some sort of a narrow species of causation. So, if it turned out that we can control the past, this would be even more surprising than if we could merely cause it.

Back in section 3, we saw that Albert defended his account by claiming backward causation is unverifiable. However, some have countered this by describing scenarios where *our awareness* of this phenomenon is a built-in feature. They then go on to argue that in such circumstances, we can not only cause past events but also *control* them. In this section, I shall use the dispersion mechanism sketched earlier to argue that in these situations, backward causation probably isn't happening in the first place, so these alleged instances of backward control fall at the first hurdle.

Our first task is to pin down what it might take to control an event. There are few who doubt that causation is a necessary condition for control. The sufficient conditions, however, are less clear-cut. According to Frisch (2010), these amount to

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

"causation plus epistemic access to the consequences." According to Albert (2014) and Fernandes (forthcoming), these amount to something stronger: "causation plus profitability of the consequences." The two examples that follow purport to satisfy these two sets of sufficient conditions (respectively) and thereby show how Albert's account allows situations where we can control the past. Let's look at these in turn.

The first case, from Frisch (2010), runs as follows: Suppose I'm playing the piano, and a certain melody appears in the score twice, where it's followed by two different endings. So, the score proceeds as follows: melody, first ending, melody, second ending. Now, during my performance, I can't remember whether I played the melody once or twice. However, I know from experience that I always play this song correctly. This means that when I play the first ending, this is a reliable record of my having played the melody once, and when I play the second ending, this is a reliable record of my having played the melody twice. Let's also assume that which ending I decide to play is the *only* record of how many times I played the melody.

For argument's sake, imagine I play the first ending (x). This makes it highly likely that I played the melody just once (y), and so $\Pr(y \mid N_0) \approx 1$. x causes y iff the causal criteria are satisfied, and because Criterion 1 is probable, the key question is whether or not Criterion 2 is probable. But this *is* probable, for had I played the second ending ($\neg x$), it would be highly likely that I played the melody twice ($\neg y$), and so $\Pr(y \mid N_\star) \approx 0$. Because the causal criteria are probably satisfied, x probably causes y. Because there are no records relevant to y besides the antecedent itself, Albert's barrier to backward causation is overcome. Moreover, unlike in the Atlantis or carp scenarios, it's part and parcel of the situation that I'm aware of this probable causal influence. This is because the records responsible for causal influence are my own actions, and I'm aware of when I perform these. If the sufficient conditions for control are "causation plus epistemic access to the consequences," then my decision to play the first or second ending in the present allegedly controls how many times I played the melody in the past.

Albert's (2014) response to this example is that even if we have epistemic access to the consequences of our actions, this doesn't count as control because we can't profit from this knowledge. This brings us to Fernandes, (forthcoming) example, which attempts to satisfy this stronger condition. I lay this out next.

Imagine I am seated in a room that has a persistent fly buzzing around it, and there is a video camera that records the fly's location at t_1. Additionally, this video camera is connected to a screen that reveals the fly's location a moment later at t_2. This apparatus was rigged up by a sadistic scientist who, upon observing the screen at t_2, gives me a reward iff the fly buzzes past my face at t_1. But as it happens, I am an expert flyswatter, which means my swatting at t_2 is a reliable record of a fly buzzing past my face at t_1, and my remaining still at t_2 is a reliable record of a fly *not* buzzing past my face at t_1. In this scenario, the only records of the fly's location at t_1 are my behavior and the image on the screen.

Now, let's suppose that a fly buzzes past my face (y) at t_1, and I swat at it (x) at t_2. Does x cause y? Criterion 1 is satisfied by hypothesis. To probe Criterion 2, we consider the implications of me not swatting ($\neg x$) at t_2: in all likelihood, this would mean the fly *didn't* buzz past my face ($\neg y$) at t_1, so $\Pr(y \mid N_\star) \approx 0$. As in Frisch's example, both causal criteria are probably satisfied, so my swatting probably causes the fly to buzz past my face. But this time, whether or not I swat has an extra implication: it

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

determines the image on the screen, and hence it determines whether or not I get a reward. As in the piano case, my awareness of backward causation is built into the scenario. But this time, my action ($\neg x$) causes *other* present records of whether or not y occurred, from which I may profit. So, if the sufficient conditions for control are "causation plus profitability of the consequences," then my decision to swat or not swat in the present allegedly controls the fly's location in the past.

Most discussions about whether or not these scenarios involve backward control relate to the sufficient conditions: what these are and whether or not they're satisfied.[13] It's generally taken for granted that backward causation—a necessary condition—is in play. However, I shall deny this basic premise by appealing to the dispersion mechanism. This means we can discount these two alleged instances of backward causation while sidestepping this more in-depth debate about the sufficient conditions for control.

The previous two examples share a key similarity: backward causation is allegedly in play because N_0 contains no records of the consequent that also feature in $N_\star$ (in the piano case, this is because the only record is my action; in the fly case, this is because the only records are my action plus other records it causes). If such records existed, they would cement the matter of whether or not y occurred, preventing its probability from varying much between N_0 and $N_\star$ and thus obstructing backward causation/control.

However, precisely such records generally *do* exist in the present. Most significantly, if my present actions are genuine records of what I did a few moments ago, then these cannot simply pop out of nowhere. They must have physical precursors that have persisted since that time—presumably, records that have been stored in my brain. But the dispersion mechanism suggests that there will probably be many other records at this time: light leaving the window, air currents, DNA traces, and so on. These are often relatively motley and unreliable, hence why we tend to ignore them when envisaging scenarios like the ones described earlier. But they probably exist all the same. Therefore, whether or not I play the second ending or swat will probably *not* cause past events because this will be fixed (with high likelihood) by these many records in the present. So, backward causation/control is unlikely in both scenarios.

It is tempting to respond as follows: When evaluating counterfactuals, the explicitly stated antecedent $\neg x$ isn't the *only* respect in which $N_\star$ differs from N_0. We almost always have to flesh out $N_\star$ with additional alterations. At the bare minimum, some gravitational fields must be redrawn to account for the fact that the antecedent entails a different matter distribution from our world—however slight.[14] Therefore, perhaps $N_\star$ is not just a world in which I *act* differently in the present but also a world in which all records—including those in my brain—indicate a different past than the one indicated by N_0.

The problem, however, is that this is asking for an awfully large gulf between N_0 and $N_\star$. In asking for all records of my melody-playing or the fly's location to be different, we're not just asking for a world in which I act differently in the present. We're

[13] For instance see Loewer (2012), Albert (2014), and Loew (2017).

[14] It's on these sorts of grounds that Fernandes argues that had I not swatted, the screen would display a different image.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

also asking for a world in which my brain's contents are different, the light that left my window is different, the air currents are different, and so on. So, even if we grant that antecedents bring with them some extra changes (e.g., redrawn gravitational fields), a world *this* different from our own seems like a poor candidate for qualifying as $N_{\star}$.[15]

Of course, one is free to simply stipulate a scenario in which the dispersion mechanism fails and there are genuinely no other records besides the ones Frisch and Fernandes describe. In these highly unusual circumstances, backward causation would indeed be likely, and my proposed obstacle to backward control would be lifted. However, we generally wouldn't know whether or not this situation obtains because the sorts of records I've been describing are motley and unreliable. So once again, we run into a familiar distinction: even when the chance of backward causation is high (as in this instance), we'd be none the wiser, so Albert's original unverifiability defense rearises.

7. Conclusion

Albert's account allows backward causation whenever an unrecorded event lies in our past. One well-known objection is that we never observe backward causation. However, Albert's response is that this phenomenon is by its nature unverifiable, hence why we never seem to encounter it.

When we combine this account with the commonsense idea that minor events lack records in their distant futures, the result is that backward causation is ubiquitous. This is an unattractive feature of Albert's causal account when we read it at face value. To remedy this, I have proposed the dispersion mechanism, a process in which records seem to vanish while in fact proliferating.

This analysis also defends Albert's account from a second well-known objection, which is that it allows backward control. Although the sufficient conditions for this are much debated, it is widely accepted that backward causation is a necessary condition. Because my proposal obstructs this precondition, purported cases of backward control are nipped in the bud.

References

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

Albert, David. 2014. "The Sharpness of the Distinction between the Past and the Future." In *Chance and Temporal Asymmetry*, edited by A. Wilson, 159–74. Oxford: Oxford University Press.

Arntzenius, Frank. 1990. "Physics and Common Causes." *Synthese* 82 (1):77–96.

Dummett, Michael. 1978. "Bringing About the Past." In *Truth and Other Enigmas*, 333–350. Cambridge, MA: Harvard University Press.

Fernandes, Alison. Forthcoming. "Time, Flies, and Why We Can't Control the Past." In *Time's Arrows and the Probability Structure of the World*, edited by B. Loewer, E. Winsberg, and B. Weslake. Cambridge, MA: Harvard University Press.

Fine, Kit. 1975. "Critical Notice: Lewis, *Counterfactuals*." *Mind* 84 (335):451–58.

Frisch, Mathias. 2005. *Inconsistency, Asymmetry, and Non-Locality: A Philosophical Investigation of Classical Electrodynamics*. Oxford: Oxford University Press.

[15] This is analogous to how Lewis (1979) defuses Fine's (1975) objection about Nixon and the button.

https://doi.org/10.1017/psa.2022.14 Published online by Cambridge University Press

Frisch, Mathias. 2007. "Causation, Counterfactuals, and Entropy." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by H. Price and R. Corry, 351–95. Oxford: Oxford University Press.

Frisch, Mathias. 2010. "Does a Low-Entropy Constraint Prevent Us from Influencing the Past?" In *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics*, edited by A. Hüttemann and G. Ernst, 13–33. Cambridge: Cambridge University Press.

Frisch, Mathias. 2014. *Causal Reasoning in Physics*. Cambridge: Cambridge University Press.

Hemmo, Meir, and Orly R. Shenker. 2016. *Maxwell's Demon*. Oxford Handbooks Online. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199935314.013.63.

Kim, Jaegwon. 1973. "Causes and Counterfactuals." *Journal of Philosophy* 70 (17):570–72.

Kutach, Douglas N. 2002. "The Entropy Theory of Counterfactuals." *British Journal for the Philosophy of Science* 69 (1):82–104.

Lewis, David. 1973a. "Causation." *Journal of Philosophy* 70 (17):556–67.

Lewis, David. 1973b. *Counterfactuals*. Cambridge, MA: Harvard University Press.

Lewis, David. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs* 13 (4):455–67.

Lloyd, Seth. 1994. "Causal Asymmetries from Statistics." In *Physical Origins of Time Asymmetry*, edited by J. J. Halliwell, J. Pérez-Mercader, and W. Zurek, 108–16. Cambridge: Cambridge University Press.

Loew, Christian. 2017. "The Asymmetry of Counterfactual Dependence." *Philosophy of Science* 84 (3): 436–55.

Loewer, Barry. 2007. "Counterfactuals and the Second Law." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by H. Price and R. Corry, 293–326. New York: Oxford University Press.

Loewer, Barry. 2011. "The Emergence of Time's Arrows and Special Science Laws from Physics." *Interface Focus* 2 (1):13–19.

Loewer, Barry. 2012. "Two Accounts of Laws and Time." *Philosophical Studies* 160 (1):115–137.

Paul, Laurie Ann, and Ned Hall. 2013. *Causation: A User's Guide*. Oxford: Oxford University Press.

Reiss, Julian. 2015. *Causation, Evidence, and Inference*. New York: Routledge.

Cite this article: Stradis, Athamos. 2023. "Backward Causation: Harder Than It Looks." *Philosophy of Science* 90 (1):77–91. https://doi.org/10.1017/psa.2022.14