LSE CPNSS

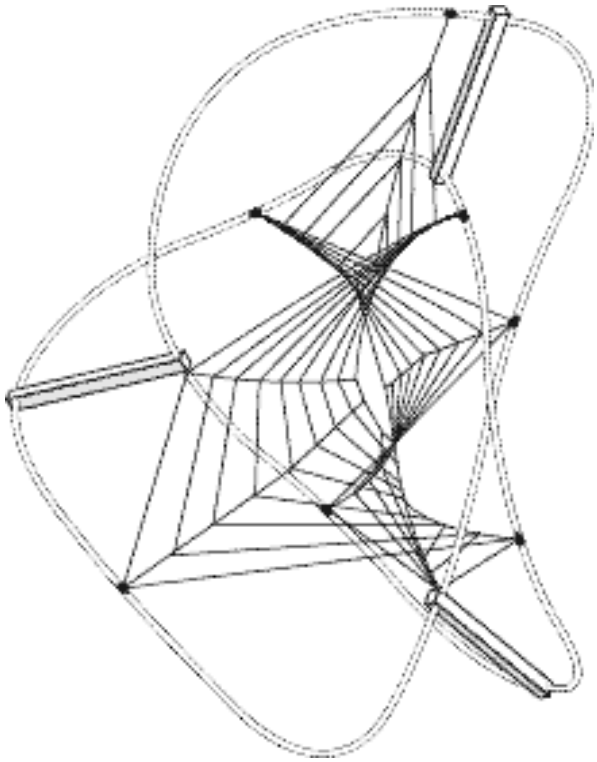
Centre for Philosophy of Natural and Social Science

Causality: Metaphysics and Methods

Technical Report 16/04

Causation, Counterfactuals and the Past-Hypothesis

Mathias Frisch



Editor: Julian Reiss

Causation, Counterfactuals and the Past-Hypothesis

Mathias Frisch
Department of Philosophy
University of Maryland
Skinner Building
College Park, MD 20742
mfrisch@umd.edu

April 2004

1 Introduction

Bertrand Russell famously argued that the notion of cause has no place in modern fundamental physics, where it has been replaced by the concept of functional dependency (Russell 1918). Fundamental dynamical laws specify how one thing follows after another, but since these laws are time-symmetric, they do not support an asymmetric distinction between cause and effect of the kind that appears to be part of our common sense notion of cause. In a recent paper Hartry Field has endorsed Russell's conclusion but has pointed to a problem resulting from Russell's thesis (Field forthcoming). Even if we were convinced by Russell's thesis, we cannot simply excise any 'weighty' asymmetric notion of cause from our conception of the world, since, as Nancy Cartwright has argued (Cartwright 1979), just such a notion seems to be essential for the distinction between effective and ineffective strategies: In deliberating which actions will further our goals we need to appeal to a robust distinction between causes and their effects, for, intuitively, we can influence the occurrence of an event by affecting the occurrence of its causes but not by influencing its effects. In fact, Field maintains that trying to reconcile the apparent need for causation in a theory of effective strategies with Russell's thesis is "the central problem in the metaphysics of causation." What, then, is the source of the causal asymmetry? How do we locate causation within a fundamental universal physics with time-symmetric laws? And if invoking the concept of cause in fundamental physics in fact proved to be a "relic of a bygone age," how is it that we have come to understand the world in asymmetric causal terms?

In this paper I want to examine one recent attempt at providing answers to these questions: David Albert's entropy-account of causal influence (Albert 2000, ch. 6).[1] Albert appears to agree with Russell's conclusion that the fundamental dynamical laws of physics neither support nor require a 'weighty' concept of cause, yet he argues that the fact that the universe had an extremely low-entropy past can account for our possession of such a concept. That is, according to Albert, the causal asymmetry can be explained in terms of an asymmetry of physical *initial* or *boundary conditions*. Yet Albert denies that there is a completely general asymmetry of causation or causal influence. Rather, he claims that the causal asymmetry is grounded in a counterfactual asymmetry that exists for small yet macroscopic hypothetical interventions of the kind for which we humans could in principle be responsible. Thus, if Albert were right, there is no tension between fundamental physics and the demands of a theory of effective strategies despite the fact that the micro-dynamical laws *alone* do not support an asymmetric notion of cause. For the causal asymmetry would turn out to be due to an asymmetry of initial conditions and not due to an asymmetry of the dynamical laws; and the asymmetry would arise precisely for those kind of possible macro-interventions that play a role in a theory of effective strategies. Field's problem would be solved.

Similar to David Lewis (Lewis 1986a, 1986b), Albert advocates a counterfactual analysis of causation. Thus, at the heart of his account is an argument for a thermodynamic entropy-account of an asymmetry of counterfactuals, which in turn is meant to explain the asymmetry of causation. Yet Albert's account has the advantage over Lewis's that it does not invoke the dubious notion of miracles that can be ranked with respect to their sizes. Moreover, while Albert

[1] Interestingly, Field himself also suggests that the solution to the problem may lie in recognizing the importance of statistical regularities to the concept of causation, similar to those that arguably can account for the concepts of entropy and temperature. As Field points out, such a solution to the metaphysical problem has the consequence that the causal asymmetries is absent on the micro-physical level.

adopts Lewis's core idea that facts about the present allow us to make inferences about the past in a way different from inferences about the future – Lewis's notion of *postdeterminants* is echoed in Albert's notion of *records* – his account does not rely on the problematic (and in fact provably false!) thesis of an asymmetry of overdetermination between the past and the future.[2] Thus, Albert's account might be understood as offering a defense of Lewis's overall project that avoids some of the latter's deep problems.

In this paper I want to argue, however, that Albert's thermodynamic account of the causal and counterfactual asymmetries is problematic as well. In the next section I will make some introductory remarks concerning the putative time-asymmetries of counterfactuals and of causation. Then I will present what I take Albert's account of the time-asymmetry of counterfactuals to be. I will criticize the account in section four, where I argue that Albert's account relies on several unargued for assumptions and has a number of highly counterintuitive consequences and, hence, ought to be rejected. I will end with a brief conclusion.

2 Counterfactuals and Causes

Albert, following Lewis, posits as a central *explanandum* that counterfactuals exhibit a time-asymmetry: In certain standard contexts the future counterfactually depends on the past, but the past does not counterfactually depend on the future. Yet one might appeal to considerations similar to those advanced by Russell to argue that all scientifically respectable counterfactuals are time-symmetric. Take a physical theory with time-symmetric dynamical laws that pose a well-defined initial value-problem. Then we can both predict and retrodict the evolution of a system governed by that theory, given the state of the system at a certain time. Moreover, the laws do not only allow us to derive the evolution of an actual system, but also allow us to determine how the evolution would have been different had the system's 'initial' state been different, where it makes no difference whether the 'initial' state occurs before or after the state in which we are interested. Thus, the laws seem to support both forward-looking and backtracking counterfactuals equally: If the state of the system at the initial time were different, both its past and its future would have to be different. Just as the future counterfactually depends on the present initial state, so apparently does the past.

Now, both Lewis and Albert believe that at least in worlds as complex as ours we do not evaluate the truth of a counterfactual by simply letting the relevant counterfactual present state of the world evolve in accord with the dynamical laws. Lewis appeals to a complicated similarity metric between worlds and maintains that the closest counterfactual worlds to ours are 'miracle worlds' diverging from the actual world, in which the laws of the actual world are not exceptionless truths; while Albert argues, as we shall see in more detail shortly, that our inferences about the past are also constrained by the hypothesis of a low-entropy past. Yet there clearly are standard scientific contexts in which we draw inferences about states of a system at different times but where our reasoning does not presuppose a rich and complex world and not one with thermodynamic features. In those contexts we draw inferences based on special, highly idealized circumstances in which the system in question can be represented as relatively simple, perhaps purely macroscopic system – for example, as mechanical or electromagnetic system. And, *pace* Lewis and Albert, the appropriate procedure for drawing counterfactual inferences in such cases can simply be to investigate the evolution of possible states of the system, given

[2] For criticisms of Lewis's account see (Elga 2001), (Field forthcoming), and (Frisch forthcoming).

certain initial or final values and the relevant dynamical laws. For example, in the context of examining possible trajectories of balls on a billiard table, it might be correct to assert the backtracking counterfactual that if a certain ball had gone into the corner pocket, then it would have been struck differently from the way it actually has been struck; just as it might be the correct thing to say that if the ball were struck differently, then it would roll into the corner pocket.

Thus, there certainly appear to be contexts for evaluating counterfactuals in which forward looking counterfactuals are not privileged. Nevertheless, I think that Lewis and Albert are correct in claiming that there *also* is a sense in which we think that the future but not the past depends on the present and, hence, that there also are contexts in which counterfactuals are time-asymmetric.

We appear to take the past to be counterfactually independent of the present in contexts which we intuitively think of as causal. On this point, I think, there is relatively widespread agreement, even though accounts of the precise relation between causal and counterfactual claims differ widely. One way to spell out the connection between causation and asymmetric counterfactuals is in terms of the notion of hypothetical interventions: Interventions into a cause influence the occurrence of its effects, but intervening into a putative effect cannot influence the occurrence of its causes (see for example Hausman 1998; Woodward 2003). One might then try to distinguish between asymmetric, intuitively causal contexts, on the one hand, and symmetric contexts, on the other, by invoking a distinction between *closed* and *partially open* systems. Counterfactuals associated with closed systems appear to be symmetric: Each set of possible initial conditions at a time defines a different closed system whose past and future evolution is given by the relevant dynamical laws. Systems with different initial states will in general have both different futures and different pasts. By contrast, counterfactuals associated with interventions from the outside into an otherwise closed systems might be thought to be asymmetric, since interventions may be taken to affect only the future evolution of the system but not its past.[3]

Two aspects of this scheme are worth being made explicit. First, the scheme is most naturally spelled out as not advancing an account of the causal asymmetry. *Intervention* arguably is itself a causal and the account simply stipulates that interventions influence the causal 'future' of a system and not its causal 'past'. Second, it appears to be crucial to this way of thinking about causation that causal systems have an environment or 'outside' from which interventions can occur. This may suggest the Russellian claim that there is indeed no room for the notion of cause in a *universal* physics that aims to have models of the universe as a whole among its class of models.[4]

In sharp contrast with this scheme, Albert proposes an account of 'intervention' that applies to closed systems as well, thereby promising to provide a place for causal notions even within a universal physics. Hypothetical 'interventions,' on Albert's account, are treated by postulating counterfactual initial states, where both future and past evolutions of the system in question are then determined by our 'normal procedures of inference,' which include use of the fundamental dynamics, but do not invoke Lewis-style miraculous violations of the laws. Instead of appealing to a difference between open and closed systems, Albert's account suggests that the difference

[3] The notion of hypothetical interventions functions in some ways similar to Lewisian miracles, with the advantage that an interaction of the system with its environment takes the place of counterlegal time-evolutions.
[4] See, for example, (Hausman 1998).

between symmetric and asymmetric counterfactuals is due to a difference between thermodynamic systems and non-thermodynamic systems. In the former case, Albert argues, a counterfactual present state that differs only locally from the actual present will be overwhelmingly likely to have evolved from a past identical to the actual past.

3 Albert's Argument

Albert begins his discussion by introducing the notion of a "causal handle". Now, in light of the preceding remarks it may come as somewhat a surprise that Albert suggests that one can introduce this notion even in the absence of any thermodynamic considerations. Independently of any appeal to a low-entropy past Albert points out that if we constrain the *remote* past of a system, then only very special alterations of the present can lead to a different *recent* past, while many alterations of the present may lead to a different future. Albert asks us to consider a collection of billiard balls such that ball 5 is currently stationary with the additional constraint that ball 5 was moving 10 seconds ago. Given the additional constraint, that ball 5 has been involved in a collision in the past 10 seconds is determined by facts about the present state of ball 5 *alone*. That is, alterations to the present state of the balls *not* involving changes in the state of ball 5 cannot change that ball five was involved in a collision during the last 10 seconds. (In fact, there will be many alterations to the present not involving ball 5 that will result in a present state inconsistent with the additional constraint.) From this Albert concludes that there are a far wider variety of "*causal handles* on the future of the ball in question here, under these circumstances, than there are on its past." (Albert 2000, 128)

An obvious objection at this point is that evaluating counterfactual situations compatible with a seemingly *ad hoc* time-asymmetric constraint on the past but not the future tells us nothing about the causal structure of the case and, in particular, cannot license any conclusions about an asymmetry of causal influence. Why are "these circumstances" the right circumstances for assessing the causal structure? The asymmetry obviously is the result of imposing an asymmetric constraint on possible alterations. If instead we were only interested in possible changes to the present that are compatible with a constraint on the *future* evolution of the system without constraining the past, then many more backtracking than forward directed counterfactuals would presumably come out true. Why then should we impose a constraint only on the past? One answer might be that the past is fixed while the future is open. Alternatively, we could maintain that we ought to keep the causal history of the event fixed, but not its future effects. But obviously these answers would beg the question, if our aim is to give a counterfactual analysis of the notion of causal influence.

It seems to me that in *non-thermodynamic* systems, such as an idealized collection of billiard balls, there is no motivation for a time-asymmetric constraint short of appealing to already explicitly causal assumptions. But in the end Albert is not interested in this case. Rather, Albert's ultimate aim is to argue that the counterfactual asymmetry arises in systems that are complex enough to exhibit thermodynamic features. Thus, he himself would probably agree that introducing the notion of a "causal handle" simply by reflecting on the consequences of imposing a time-asymmetric constraint on the motion of the billiard balls is somewhat misleading.

The new ingredient in the case of thermodynamic systems is that there is a special asymmetric constraint on the past – the hypothesis of an extremely low entropy initial state. This condition, which Albert calls "the *past-hypothesis*", is a central assumption in standard accounts

of the thermodynamic asymmetry that the entropy of a closed system never decreases. There it is needed to avoid the *reversibility objection* against the most straightforward attempt of accounting for the increase in entropy. While one can argue that the entropy of a given low-entropy macro-state increases, assuming an intuitively plausible probability distribution over micro-states and a Newtonian time-symmetric micro-dynamics, the same type of argument also allows us to conclude wrongly that the present macro-state is at a local entropy minimum and evolved from a higher entropy past. The undesirable retrodiction that entropy decreased in the past can be blocked, if the distribution of micro-states is conditionalized not only on the present macro-state but also on a low-entropy past and, ultimately, a low-entropy initial state of the universe.

Now, why should we keep the past-hypothesis fixed in assessing counterfactual changes to the present? How is this hypothesis different from the apparently question-begging assumption to hold the past state of billiard ball 5 fixed? We may, as Albert claims, have good inductive reasons to believe that the past-hypothesis is satisfied in the actual world. Yet there is much that we know about the future in the actual world that we do not keep fixed in assessing counterfactual changes to the present. Albert's answer is that in assessing the truth of counterfactuals we need to consider other worlds that are in important ways like ours. In particular, the counterfactual worlds to which we appeal in assessing the results of counterfactual changes to the present have to license the same *normal procedures of inference* as the actual world does. And Albert argues that these procedures rely crucially on assuming the truth of the past-hypothesis. If all counterfactual reasoning must rely on our normal procedures of inference from the state of the world at one time to the state at other times and these procedures presuppose the past-hypothesis, then all counterfactual reasoning must presuppose the past-hypothesis.

In somewhat more detail, Albert's argues the following.[5] In order to assess the truth of a counterfactual, where the antecedent involves some small yet macroscopic alteration to the actual world, we need to look at counterfactual worlds that are like the actual world except for the small change and then use our normal procedures of inference to determine the past and future evolutions of such worlds. In the case of forward looking counterfactuals these procedures amount to taking the present macro-state of the world, assuming an equi-probability postulate over micro-states compatible with that macro-state and evolving the state forward in accord with the dynamical laws. Thus, counterfactuals such as 'If I flipped the light switch the light would go on' come out true (assuming that in the actual world the light is and remains off

[5] I want to distinguish Albert's account from another "entropy account of counterfactuals." Douglas Kutach has recently proposed (and critically assessed) one version of such an account (Kutach 2002). Kutach cites Albert's discussion as one of his inspirations, yet there are important differences between the two accounts. For one, Kutach's account does not appeal to a notion analogous to Lewis's postdeterminants or Albert's records. In fact, Kutach does not agree with Albert that counterfactuals that are concerned with assessing the effect of local changes to the present are always evaluated under the assumption that the present is held strictly fixed aside from the local change at issue. Instead Kutach argues that there is a condition that he calls "mundaneness" restricting the range of relevant counterfactual worlds that conflicts with a condition he calls "locality" and that may require us to consider counterfactual worlds which differ from the actual world in more than the small, local change at issue. Moreover, contrary to both Lewis and Albert, Kutach believes that certain backtracking counterfactuals postulating a dependence of the past of the present are true. Rather Kutach rejects the idea that counterfactuals are evaluated by keeping the present *strictly* fixed aside from the local change at issue, precisely because he takes this procedure to conflict with the truth of certain in his mind plausible backtracking counterfactuals. Consequently, Kutach also does not believe that his theory can provide a counterfactual analysis of the notion of causal influence. My focus here will be exclusively on Albert's account.

and I don't flip the switch), if most micro-states compatible with the initial macro-state evolve into micro-states corresponding to a macro-state such that the light is on.[6]

However, backtracking counterfactuals such as 'If I had flipped the light switch, the light would have to have been on prior to that' are not supported by our normal procedures of inference.[7] If we simply evolved the counterfactual present state backward in accord with the macroscopic regularities with which we are familiar, then the light's being off after the light switch was flipped presumably would have to have been preceded by the light's being on originally. For generally flipping the switch is accompanied by a change in the light's state from on to off or vice versa. But Albert argues that this inference would violate the presupposition that there are *records* of the past. In our example, records of the light's being off in the actual world might include the fact that there is no light that has been escaping through the window a short while ago and that the light bulb is relatively cold. Albert's notion of record plays a role analogous to that of Lewis's postdeterminants: a record is a relatively localized fact about the present from which we can infer the occurrence of some event in the past. If the present contains a record of the light's having been off, then we are licensed to infer that the light was in fact off. Like Lewis, Albert believes that we take many of the local facts we do hold fixed to be sufficient for the occurrence of certain (relatively localized) events in the past. Yet, unlike Lewis, Albert is well aware that there are no local facts about the present which *alone* are sufficient for the occurrence of some past event. Rather, inferences appealing to records, Albert holds, are always inferences from facts at two different times to a fact at a third time in between, since facts cannot function as records of the past unless we assume something about the more remote past that functions as "ready condition".

Recall Albert's example of the collection of billiard balls. Albert points out that the fact that ball 5 is currently stationary functions as a record of the fact that the ball underwent a collision in the last 10 seconds *given* the additional constraint that ball 5 was moving 10 seconds ago. In other words, that the ball was moving 10 seconds ago functions as "ready condition" allowing us to record the ball's collisions in terms of its present state of motion. Without the ready condition we could not infer whether ball 5 underwent a collision from the current state of ball 5 alone but would need to know the present state of the entire collection of balls.

What does all this have to do with the past-hypothesis? Ultimately, Albert claims, the single assumption that on its own can ensure that we can treat facts about the present as records of the

---

[6] Note that given his appeal to procedures of inference, one might think that, unlike Lewis, Albert is offering *assertibility* conditions of counterfactuals and not *truth* conditions. Yet Albert concludes his discussion of counterfactuals by saying "And it follows – if all this is right – that the future does indeed counterfactually depend on what we do now, and the past […] does not." (130) This suggests that Albert is committed to the view that whatever follows from our normal procedures of inference is true.

[7] This example is not Albert's own, who also follows Lewis in trying to stack the deck through his choice of examples. Albert considers the forward looking counterfactual 'If the president pushed the button, there would be a nuclear explosion' and contrasts it with the backtracking counterfactual 'If the president pushed the button, then there would have been an explosion'. He says that there "are (for example) no worlds *at all*, even *remotely* like our own, in which the [normal procedures of inference] translate small hypothetical present differences in the present position of anybody's finger into differences between a certain thermonuclear device's exploding or not exploding two minutes *ago*." (Albert 2000, 129-30). While it might well be that this particular backtracking counterfactuals comes out false, this of course does not show that backtracking counterfactuals are false in general. The passage continues as follows: "And that (as I said before) is precisely because there is a past-hypothesis and not a future one. That (to put it another way) is because there are – vis-à-vis such things as the *past* explosion of thermonuclear devices (or the lack of them) – such things as *records*, as *memories*." (*Ibid*., all italics in original) I will criticize equating the truth of the Past Hypothesis with the existence of records below.

past is the past-hypothesis. Treating a fact as record presupposes that some ready condition in the more remote past was satisfied. But how do we know the latter fact? Again, we need some record of the ready condition's being satisfied, which in turn requires an even earlier ready condition. According to Albert, this regress ends with the past hypothesis, which can function as a first "mother of all ready conditions." That is, Albert's amended version of Lewis's thesis of overdetermination of the past by the present is that *given the past-hypothesis* localized facts about the present are records of the occurrence of certain localized facts in the past. Thus, while the light bulb's being cold alone is not sufficient for the light's having been out, it is sufficient, or at least is overwhelmingly probable, on Albert's view, in conjunction with the assumption that our universe had an extremely low entropy past.

The broad structure of Albert's argument, then, is this. Albert argues for two claims:

(*i*) If the past-hypothesis is true, then there are records of the past.
(*ii*) If there are records of the past, then there is no counterfactual dependence of the past on the present.

From (*i*) and (*ii*) Albert's conclusion follows:

(*iii*) If the past-hypothesis is true, then there is no counterfactual dependence of the past on the present.

At the heart of the account is the idea that the past counterfactually depends on the present, if there is an actual present event *c* and actual past event *e* such that our normal procedures of inference license us to accept the counterfactual 'If *c* had not occurred *e* would not have occurred.' This raises the question as to how exactly according to our normal procedures of inference we ought to evaluate the truth of counterfactuals. Yet given the question's central importance to Albert's account, it is surprisingly difficult to find a precise statement of Albert's answer to this question. The most plausible proposal, which has been implicit in some of what I said above, appears to be that our normal procedures of inference involve calculating conditional probabilities. The truth of the counterfactual 'If *c* had not occurred, *e* would not have occurred' is assessed by looking at the class of worlds that satisfy the past-hypothesis and whose present macro-states match the macro-state of the actual world as much as possible, given that *c* does not occur in those worlds. The past and future of these worlds is then determined by the micro-dynamical laws. If in most of these non-*c* worlds *e* does not occur, then the counterfactual is true, otherwise it is false.[8]

Thus, on what I take to be the most plausible reconstruction of Albert's account, if the past does not counterfactually depend on the present, then for all (suitably localized) actual present macro-events *c* and for all past macro-events *e* the following condition is satisfied: The conditional probability of *e*'s not occurring is extremely low, *given* the dynamical laws, the past-hypothesis, and that the present is unchanged except for *c*'s not occurring;[9] that is, if the past

[8] A similar scheme for assessing assertibility conditions for counterfactuals is advocated in (Kutach 2002).
[9] One might worry that since conditional probabilities come in degrees it is not clear how this can result in conditions for the truth or falsity of backtracking counterfactuals. But since the relevant thermodynamic probabilities are usually either absurdly small or very, very close to one, this might perhaps license Albert's conclusion that "the future does indeed counterfactually depend on what we do now, and the past […] does not."

does not counterfactually depend on the present, then

$$\Pr(\sim e/\sim c \& S_a \& PH) \approx 0 \qquad (1)$$

for all actual events c and e such that e is in the past of c, where PH is the past-hypothesis. Here I am taking events to be the goings on in some region of space at a particular time. The event c is the complete actual present macro-state in some suitably small region of space. S_a is the remainder of the present macro-state of the world. Thus, $c \& S_a$ is the complete present macro-state of the world. Also I am assuming that $\sim c \& S_a$ is nomically possible – that is, that the occurrence of c is not implied by S_a together with synchronic constraints imposed by the laws.

4 Criticism

4.1. Records and the Past-Hypothesis

Albert maintains that local facts about the present can function as records of the past, if we can assume certain facts about the remote past as "ready condition," and that "the initial macro-condition of the universe as a whole" can function as the "mother of all ready conditions." (118) From this he immediately and without further argument concludes the following:

> And so it turns out that precisely the thing that makes it the case that the secondary law of thermodynamics is (statistically) true throughout the entire history of the world is also the thing that makes it the case that we can have epistemic access to the past which is not of a predictive/retrodictive sort; the reason there can be records of the past and not the future is nothing other than that it seems to us that our experience is confirmatory of a past-hypothesis but not of a future one. (Albert 2000, 118)

And a little further on he says:

> [E]verything we can know of the past and present and future history of the world can be deduced, in its entirety […] from the following four elements: what we know of the world's present macrocondition […]; the standard microstatistical rule; the dynamical equations of motion; the past-hypothesis. (*Ibid.*, 119)

Hence, Albert takes the claim that "the initial macro-condition of the universe as a whole" functions as ready condition to imply premise (*i*) – the claim that the past-hypothesis can play the role of "mother of all ready conditions."

Clearly Albert's account relies on the claim that the past-hypothesis is sufficient as ready condition. For if instead the complete initial macro-state of the universe was required as ultimate ready condition, then our practice of appealing to records would remain utterly mysterious, since the complete initial macro-state is unknown to us. Yet by the same token, to conclude from the claim that the initial macro-condition of the universe can function as ready condition that the past-hypothesis alone is such a ready condition is a *non-sequitur*. For the past-hypothesis provides us with significantly less than a full specification of the initial macro-state of the universe. All the past-hypothesis asserts is "that the world first came into being in whatever particular low-entropy highly condensed big-bang sort of macro-condition it is that the normal inferential procedures of cosmology will eventually present to us." (96) And clearly whatever it is that cosmology will eventually present us with, this will fall far short of a complete account of the initial macro-state of the universe. Thus, Albert owes us an argument for why the broad constraints on the early universe posited by the past-hypothesis (as opposed to a full specification of the macro-state of the early universe) are sufficient to ensure that local facts about the present

can function as records of the past.

To illustrate this point, we might imagine a slightly amended version of the example of the billiard balls. Let us assume that ball 5 is currently *moving* and was *stationary* 10 seconds ago. Further, let us imagine that the balls are moving on a table with weak frictional forces. Given the ready condition that ball 5 was stationary, the fact that the ball is currently moving functions as a record of a collision in the last 10 seconds. The ready condition in this case exactly specifies the value of one of the system's state-space variables. But obviously it does not follow from the fact that the ball's having been *stationary* can function as ready condition that also the claim that the system of balls was in a *low-entropy* initial state can function as ready condition. From the fact that the system of balls was in a low-entropy state we can conclude that the most likely evolution of the system of balls was one that is thermodynamically normal, and, hence, that the ball's currently moving is not due to random 'anti-frictional' forces exerted by the table. But without the further assumption that the ball was stationary 10 seconds ago, we cannot exclude the possibility that the ball has been moving without collisions for more than 10 seconds.

As far as I can tell, Albert does not offer any argument that in the case of the universe as a whole the assumption of a low-entropy past alone can function as ready condition. What Albert does argue for is the claim that there would be no reliable records in worlds that do not satisfy the past-hypothesis. In any such world, Albert says, the most probable way in which putative records originate would be as results of random fluctuations from of a maximal entropy state and, hence, they would not be correlated with the relatively low entropy states of which they are supposed to be records. But this argument can at most show that the past hypothesis is *necessary* for the existence of records, but not that it is *sufficient*, as (*i*) claims.

We can, however, try to imagine the kind of considerations that one might advance in support of premise (*i*). How, we might ask, could it be that local facts about the present are associated with a past different from that of the actual world? One way to construct such a situation is to postulate some small yet macroscopic change to the actual present and then evolve the resulting state backward in time. The effect of such local changes will in general be that *other* local facts are no longer associated with the same past events with which they were associated in the actual world: such present facts constitute fake records, as it were. In the amended billiard ball example ball 5 is presently moving in the actual world, and this is associated with the ball having undergone a collision in the past 10 seconds. But there can be changes to the state of balls *other* than ball five that, if we evolve the state of the balls backward in accord with the laws, will result in a history where ball 5 did not undergo a collision in the last 10 seconds. In the corresponding counterfactual world the fact that ball 5 is presently moving constitutes a fake record of its past evolution. Of course in such counterfactual worlds the ready condition that ball 5 was at rest 10 seconds go is not satisfied. The crucial question is whether the past-hypothesis would likewise not be true in such a world.

Adam Elga, in a somewhat different context, has presented an argument that suggests that the past-hypothesis would indeed not be satisfied in most counterfactual worlds resulting from localized macroscopic changes to the actual world (Elga 2001). Elga points out that the time-evolution of the actual world *toward the past* is thermodynamically extremely unlikely. (This is most easily seen if we imagine that the direction of time were flipped.) Moreover, the evolution toward the past is extremely sensitive to small changes in the micro-state of the world: most small changes will result in worlds that evolve in thermodynamically normal ways toward the past – that is, worlds that violate the past-hypothesis and behave anti-thermodynamically in the normal time-sense. For example, in the case of the billiard balls it is probable that changing the

position of any of the balls will, through small changes in the gravitational force, disturb the normal thermodynamic behavior of thermodynamic sub-systems in the vicinity. Further and further into the past, more and more sub-regions of such a world will be 'infected' by the anti-thermodynamic behavior with the result that the remote past of the world will have high entropy.

The upshot is that localized macroscopic changes to the present result both in an anti-thermodynamic past *and* in 'fake' records of the past. But this again is not enough to establish claim (*i*), according to which the fact that there are no localized reliable records of the past *implies* that the past-hypothesis is false. For the above construction is not the only way to generate worlds with fake records. Instead, there are also thermodynamically normal worlds in which putative records are associated with past events different from those with which they are correlated in the actual world. We can construct such worlds by postulating, as before, a small macroscopic change to the present and then evolving the instantaneous macro-state backward in accord with the *macro-dynamics* governing the world in question (instead of the underlying micro-dynamics). Often such worlds will exhibit 'strange' or inexplicable correlations, but none that can be excluded on thermodynamic grounds, since these worlds are thermodynamically normal, just like the actual world.[10]

But do the considerations above not show that such worlds are extremely improbable? The answer is 'No,' or more precisely: Such worlds are no more improbable than the actual world. Given an equi-probability distribution over micro-states compatible with the *actual* present macro-state, it is extremely improbable that that state evolved from a low entropy-past. This, after all, is just the reversibility objection that is circumvented by simply postulating a low-entropy past. That a small counterfactual macro-change to the present will again result in a micro-state which evolved from a low entropy macro-state is no less probable than the low-entropy past of the actual world.

4.2. Records in Counterfactual Worlds

The second premise of Albert's argument is the claim that if there are records of the past, then there is no counterfactual dependence of the past on the present. In discussing an example of a putative case of backward counterfactual dependence, Albert supports this claim by saying that the past could not have been different since a different past would have to have left traces or records that ought to be part of the present. Since by assumption the counterfactual present is identical to the actual present except for a small, local change, the counterfactual present contains traces of the actual past but not of any counterfactual past events. Hence, the past does not counterfactually depend on the present. This argument appears to rely on the assumption that the probability of any past event *e* given its present records *R* and the past-hypothesis *PH* is close to one:

$$\Pr(e/R\,\&PH) \approx 1, \text{ or } \Pr(e/R\,\&PH) = 1 - \varepsilon, \text{ with } \varepsilon \approx 0. \qquad (2)$$

We have just seen that the claim that the past-hypothesis alone can ensure the reliability of records is problematic, but for present purposes I want to grant that claim and see what follows from it.

[10] We might, for example, think of the wave's generated by a stone being dropped into a pond. In a counterfactual world with the same macro-dynamics that differed macroscopically only in that the stone is being dropped elsewhere the waves on the pond would appear to be due to correlated incoming waves.

According to premise (*ii*), (2) implies (1), or equivalently

$$\Pr(e/\sim c\&S_a\&PH) \approx 1. \tag{3}$$

As in the case of premise (*i*), however, Albert provides no argument for (*ii*). If (2) were a strict equality, then (3) would indeed follow and (*ii*) would not need to be introduced as an independent premise but would simply be a consequence of the probability calculus. Since, however, macro-states are only probabilistically given in terms of the underlying micro-states and their dynamics, we need a justification for the move from (2) to (3). It may be plausible that conditionalizing on the entire state S_a of which the records R are a part does not change the probability of e, *i.e.* that (2) implies

$$\Pr(e/S_a\ \&PH) = 1 - \varepsilon. \tag{4}$$

But what is less clear is why conditionalizing on $\sim c$ as well should not significantly affect the probability of e. Again, Albert owes us an argument.

There is one particular class of events for which it is perhaps most obvious that there is indeed the need for an argument here – those events e that are complete macro-states of cross sections of the backward light cone of c in the relatively recent past of c. Any such event e determines the occurrence of c with "thermodynamic certainty," as it were, if we assume that the relevant macro-laws are relativistic and near-deterministic:

$$\Pr(c/e\&PH) \approx 1 - \delta, \text{ with } \delta \approx 0. \tag{5}$$

There are, we believe, many systems like this. Billiard tables are one example. There are, of course, also systems which behave chaotically or in which the macro-dynamics is probabilistic (such as coin tosses). I am here focusing on systems which we model in terms of a deterministic macro-dynamics.

Since the probability of c is completely determined by events in its backward light cone, e will screen off c from any events outside of the light cone of c. In particular,

$$\Pr(c/e\&PH) = \Pr(c/e\&S_a\&PH)\ . \tag{6}$$

However, it follows from (3) and the definition of conditional probability that

$$\Pr(e\&\sim c/S_a\&PH) >> \Pr(\sim e\&\sim c/S_a\&PH). \tag{7}$$

That is, it is much more probable that the actual past event e occurs without the actual present than that neither the past nor the present are those of the actual world. And this is so, even though the occurrence of e makes it dynamically extremely improbable that c does not occur and there are no purely dynamic constraints that make $\sim e\&\sim c$ improbable.

Intuitively, premise (*ii*) assumes that we treat inferences based on records as more reliable than predictions and retrodictions based on the dynamics. The dynamics alone would predict that a locally different present macro-state would in general have resulted from a different past macro-state. According to (*ii*), however, any such retrodiction is overridden, as it were, by the assumption that records are reliable. Yet one might worry that this gets things backwards: No

matter how reliable our records are, they never can be more reliable – and will in general be far less reliable – than any inferences we can draw based on a complete macro-state, the past-hypothesis, and the dynamics. Of course the conjunction of (3), (5) and (6) does not contain a contradiction. Yet it is far from obvious (and requires an argument) why the record condition (2) ought to commit us to (3), and hence (7).

That Albert's account assigns a primary role to the relation between e and the record-bearing state S_a can also be seen through the following considerations. If the past is counterfactually independent of the present, then the conditional probability of any counterfactual past event e_c is much smaller than the conditional probability of the actual past e:

$$\Pr(e_c/c_c\&S_a\&PH) << \Pr(e/c_c\&S_a\&PH) \tag{8}$$

for all counterfactual events e_c and c_c such that e and e_c are in the past of c_c and e_c is a non-actual past event incompatible with e. If we once more take e to be a cross section of the past light cone of c (and e_c to be a counterfactual event in the same region), then we can derive from (8), screening-off conditions for c_c analogous to (6), and the assumption that $\Pr(e_c/PH)=\Pr(e/PH)$ the following:

$$\Pr(c_c/e_c\&PH) \,/\, \Pr(c_c/e\&PH) << \Pr(S_a/e\&PH) \;/\; \Pr(S_a/e_c\&PH).^{11} \tag{9}$$

This says intuitively that the probability of the present macro-state *outside* of the region in which c or c_c occur varies much more widely with the state of the past region in which e occurs than the probability of the state in the region of c. Whether e occurs or some counterfactual event e_c, makes a huge difference to the probability of S_a and affects the probability of c_c much less, even though e and e_c occupy a cross section of the backward light cone of c_c but of course not of S_a.

Another counterintuitive consequence of Albert's account is this. From either (3) or (4) in conjunction with (5) and (6) it follows that

$$\Pr(c/S_a\&PH) \approx 1, \tag{10}$$

which states that the macro-state in a small sub-region of the world at some time can be determined with 'thermodynamic certainty' from the macro-state of the world everywhere else at that time together with the assumption that the universe had a low entropy past. Yet in general the fundamental laws do not provide us with sufficient synchronic constraints to fix the state in a small macroscopic region given the state of the world in a large enough neighborhood of that region and it is not obvious how positing a low entropy past can provide the missing ingredient. We might think, for example, about a large number of coupled thermodynamic systems, such as a collection of gases. If the system is not in equilibrium, then knowing the state of every body of gas except one in addition to the fact that the system had a low entropy past does not allow us to infer the state of the remaining body, contrary to what (10) claims.

Finally, it follows from the record condition (4) that there is an analogous future 'record' condition as well. For consider any actual future event f. Then if the macro-dynamics is near

[11] (9) can be derived by applying Bayes's theorem to both sides of (8), appealing to the fact that cross-sections of the past light cone screen off c_c from S_a, and a second application of Bayes's theorem.

deterministic

$$\Pr(f/c\&S_a\&PH) = 1 - \gamma, \text{ with } \gamma \approx 0. \tag{11}$$

But from (11) together with (4), (5) and (6) it follows that

$$\Pr(f/S_a\ \&PH) \approx (1 - \varepsilon - \delta - \gamma)[\Pr(c/f\&\ S_a\ \&PH)\ \Pr(e/c\&S_a\&PH)]^{-1} > 1 - \varepsilon - \delta - \gamma,^{12} \tag{12}$$

and hence that

$$\Pr(f/\ S_a\ \&PH) \approx 1. \tag{13}$$

This is a future 'record' condition analogous to (4) and states that facts about the future can be determined by less than a full set of initial conditions, as would be required by the dynamics. Similar to the case of the past condition, (13) alone implies neither that the future is counterfactually independent of the present nor that it is counterfactually dependent on the present. Thus, it is open to Albert to postulate the following condition for future events *f* that are cross sections of the future light cone of *c*:

$$\Pr(f/\sim c\&S_a\&PH) \approx 0. \tag{14}$$

That is, despite the fact that there ought to be 'records' of the future as well as of the past, one can postulate that the future, but not the past counterfactually depends on the present, in accord with (3) and (14). The question, however, is what reasons we have for treating past and future differently in this respect. Intuitively, the account assumes that records ought to be weighed more heavily than inferences based purely on the local dynamics and the statistical postulate in the case of inferences from the present to the past, while the dynamics is weighed more heavily in the case of inferences to the future. But what can account for this difference? One might suggest that the difference lies in the fact that the past record condition holds independently of anything that can be derived with the help of the full dynamics, while the future record condition follows from the past record condition together with (6) and (7), which are clearly dynamical constraints. The worry about this suggestion, however, is that it might smuggle in the asymmetry. For one might have equally begun with the future record condition (13) as premise and derived (3) with the help of the dynamical constraints.

I have argued that the conjunction of Albert's entropy account of causal influence and the denial of backward counterfactual dependence has a number of highly counterintuitive consequences. Thus, if we want to retain the account but do not wish to accept the consequences, such as (10), then we appear to be forced to conclude that there is backward causal influence. At this point one might try to argue in defense of Albert's account that whatever counterfactual dependence of the past on the present there is, it is much less and dies off much faster than any dependence of the future on the present. That is, one might try to argue that neither the record condition (2) nor the counterfactual independence claims (1) hold for *all*

[12] (12) can be derived by applying Bayes's theorem to (11), obtaining an expression for $\Pr(c/S_a\&PH)$ by applying Bayes's theorem to the right-hand-side of (6) and then plugging in (5) and (4). In (12) I have only retained terms in first-order in ε, δ, and γ.

past events and that events in the very recent past of *c* will exhibit some counterfactual dependence on *c*, but that nevertheless there is a significant asymmetry in the degrees of counterfactual dependence that is sufficient to account for our asymmetric notion of causal dependence.[13]

However, this defense faces serious problems of its own. First, nothing in my criticism relies on the fact that *e* is an event in the very recent past. As long as *e* is recent enough for the macro-dynamics linking *e* and *c* to be near-deterministic, the arguments go through. Second, the defense proposes to replace what appears to be a sharp and precise distinction with a qualitative and gradual difference. According to our common sense notion of causation we think that our actions can influence the future but have *no influence at all* on the past. What a defender of an 'Albert-style' entropy-account would have to explain is how we have come to believe in this sharp distinction, despite the fact that, according to the account, there is some counterfactual dependence of the past on small interventions into the present.

A final worry is that we would still need to be given an argument for why there should be such a difference in counterfactual dependence, given the past-hypothesis. In fact, one might worry that within the general framework we ought to conclude that there is a difference of counterfactual dependence in the *opposite* direction from the one the account postulates. For, *pace* Lewis's claim that small miracles in the present would lead to ever more divergent worlds, the fact that entropy increases seems to ensure that most small differences in the present state of different possible worlds will quickly wash out. What shirt I decide to wear this morning has very little influence on the future history of the universe. The light reflected by my shirt will be absorbed by the walls of the room I am in or, if it escapes, will quickly be absorbed and randomly scattered by the earth's atmosphere. And nobody will remember the color of the shirt I wear today a week from now. Of course it is possible to construct some story according to which the future fate of the universe depends crucially on the color of my shirt today, but such scenarios are clearly not what happens normally.

By contrast, one might think that even small local differences in the present have to be associated with very large differences in the past. For either, we maintain that any small change to the present would have to be the product of an anti-thermodynamic past, which would result in a past extremely different from the actual past. Or we restrict our focus to thermodynamically normal worlds. But then small present differences would presumably have to be amplified toward the past, since the thermodynamic arrow ensures that differences in the state of a system get more and more washed out as the system approaches equilibrium. Think, for example, about a damped wave in some medium. Differences between different initial states of the wave become smaller and smaller as the medium approaches equilibrium. But, if the system is closed, then differences in the present state of the wave would have to have resulted from even larger differences in its past state.

5 Conclusion

I want to sum up what I have argued. I have offered two kinds of criticisms of Albert's account. The first kind can simply be read as an invitation to fill in some of the lacunae in the argument. As I have pointed out, Albert offers no argument for why the very general constraint given by past-hypothesis is alone sufficient as a ready condition; he provides no argument for the move

[13] This reply was suggested to me by Adam Elga and Doug Kutach.

from a record condition such as (2), to the claim that the past is counterfactually independent of the present; and, finally, there is no argument for the asymmetric treatment of past records and future 'records' despite the fact that one record condition implies the other, given a very general constraint on the dynamics.

The second kind of criticism is rather more serious and suggests that there might be no arguments forthcoming that could fill in all the gaps in the account. For I argued, first, that there are counterfactual worlds no less probable than the actual world in which present putative 'records' are associated with past events different from those with which they are associated in the actual worlds. And second, Albert's account has two consequences which we ought to reject, namely that the present macro-state of a small region of the world can be derived from the present macro-state elsewhere and that, intuitively, the causal past of such a region has a greater influence on the state of the *rest* of the world than it has on the state *in* that region. Finally, I argued that it is doubtful that Albert's account can be saved by allowing that there is a counterfactual dependence of the past on the present, albeit one that is much weaker than that of the future on the present.

Given the problems of Albert's account, I am doubtful that a convincing entropy account of the causal asymmetry can be given. Also, it seems to be well established by now that Lewis's miracle-account of the causal arrow fails as well. Even though I cannot argue for this here, I want to suggest that the proper response to these difficulties is not to try to fix the problems of counterfactual accounts of the causal asymmetry, but rather to turn Lewis's counterfactual analysis on its head: Certain counterfactuals are *causal*. These counterfactuals are asymmetric because the causal relation is asymmetric: There is no counterfactual dependence of the past on the present, because we can causally influence the future but not the past (at least for all we know). And this, it seems, has nothing whatsoever to do with the fact that entropy tends to increase. Unfortunately, Field's puzzle concerning the role of causation in fundamental physics still awaits a solution.

References

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

Cartwright, Nancy. 1979. Causal Laws and Effective Strategies. *Nous* 13:419-438.

Elga, Adam. 2001. Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science* 68 (Proceedings):S313-S324.

Field, Hartry. forthcoming. Causation in a Physical World. In *Oxford Handbook of Metaphysics*, edited by M. Loux and D. Zimmerman. Oxford: Oxford University Press.

Frisch, Mathias Florian Johannes. forthcoming. *Inconsistency, Asymmtry and Non-locality: Philosophical Issues in Classical Electrodynamics*. New York: Oxford University Press.

Hausman, Daniel M. 1998. *Causal Asymmetries*. Cambridge, U.K. ; New York: Cambridge University Press.

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

Lewis, David. 1986a. Causation. In *Philosophical Papers*. Oxford: Oxford University Press.

———. 1986b. Counterfactual Dependence and Time's Arrow. In *Philosophical Papers*. Oxford: Oxford University Press. Original edition, Nous 13 (1979).

Russell, Bertrand. 1918. On the Notion of Cause. In *Mysticism and Logic and other Essays*. New York: Longmans, Green and Co.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.