THE PHYSICAL FOUNDATIONS OF CAUSATION

DOUGLAS KUTACH

Abstract. I sketch an account of causation, showing how several of the main features of causation—asymmetry, transitivity, and necessitation (and sometimes probability-raising)—arise from the combination of fundamental dynamical laws and a special constraint on the structure of matter in the past. The necessitation and transitivity of causation are grounded in the fundamental physics, but the causal asymmetry is grounded at the coarse-grained level of macroscopic events and processes. At no single level of description does the physics justify the conditions that are taken to be constitutive of causation. Nevertheless, if we mix our reasoning about the microscopic and macroscopic descriptions, the structure provided by the dynamics and special initial conditions can justify the folk concept of causation to a significant extent, enough to explain why our causal concept works so well even though, at bottom, it is comprised of a patchwork of principles that don't mesh well.

The prominence of causation in twentieth century metaphysics is curious considering it partly stems from a classical theory of interaction by mechanical contact that has long been superceded by more sophisticated physics. Specifically, the idea that a central organizing principle of nature is a causal relation between events is not motivated by a serious examination of fundamental physics. What we do find in our best fundamental theories are equations expressing relationships between physical quantities at different times and places, equations that have no obvious connection with the concept of causation. Bertrand Russell in "On the Notion of Cause" (1913 [8]) took this observation as evidence for his argument that there is no law of cause and effect, and that causation is dispensable. While there is some justice to Russell's claim, the utility of causal notions demands an explanation. If there is just physics, why do the ideas of cause and effect serve us so well? And if there *is* a physical explanation for the usefulness of causal notions, would that not arise by demonstrating how causation reduces to the physical?

The answer to these two questions, I suggest, is that different aspects of the physics justify different principles about causation, and together these elements suffice to explain the utility of our notion of cause. Yet, the physics as a whole does not support a reduction of the robust notion of cause that philosophers usually care about, the kind of causation for example that applies to ordinary events and is useful for assigning causal responsibility and matches important pre-theoretical intuitions about causal interaction among ordinary objects.

At bottom, causation is a result of two quite different aspects of our fundamental physics: boundary conditions and dynamical laws. The dynamics vindicate our thinking of the cause as somehow necessitating the effect, and the boundary conditions vindicate our thinking of

Date: 10.15.2005.

Key words and phrases. Causation, Counterfactuals, Entropy, Past hypothesis, Common cause, Causal Asymmetry, Counterfactual Asymmetry, Backtracker.

The lion's share of improvements to this paper are due to Jonathan Schaffer's thoughtful comments. Huw Price receives my gratitude for hosting the conference on causal republicanism and generously supporting my work on this paper through the Centre for Time.

the cause as happening before the effect. Yet these two components do not cohabit peacefully. On the one hand, the necessitation relation in the physics applies only to detailed physical microstates, not to coarse-grained events. On the other hand, the causal asymmetry is grounded in the physics only insofar as one is concerned with relations among coarse-grained events, not among the detailed microstates. There is no single level of description for the relata where all the constitutive properties of causation apply. Thus, our concept of causation is a kind of arranged marriage, with the bride of necessitation and the groom of asymmetry being ill-matched but wed nevertheless for social utility.

Yet, if we allow ourselves to mix our reasoning about both microscopic and macroscopic (coarse-grained) descriptions of events and processes, we can justify a causal-like relationship that explains key facts about causation simpliciter. It explains why recourse to the idea of cause and effect is so fruitful in realistic situations. This account of causation fails to underwrite popular philosophical conceptions of causation primarily by not fully justifying intuitions about how salient causes should be distinguished. In the stock example where Billy and Suzy throw rocks at the bottle, and Suzy's rock breaks the bottle significantly before Billy's rock arrives, everyone is supposed to agree that Suzy's throw was a cause of the bottle's breaking but not Billy's. In the account presented here, both throws are counted loosely speaking as causes, but for pragmatic reasons (that are well-grounded by the physics) we attach a lot more importance to Suzy's cause than Billy's. Standard robust accounts incorporate such pragmatic features into the deep structure of causation, claiming that Billy's throw is not a cause, period. Yet the necessarily capricious selection of which events count as salient combined with the robust theorists' oversimplification by insisting that each event count as either cause or non-cause, turns out to be the source of the most serious problems for robust accounts, e.g. late preemption, double prevention, and causation by omission.

1. Causation and Physical Necessity

The 'cement of the universe' is David Hume's famous phrase describing our conviction that causes necessitate their effects. Setting aside indeterminism momentarily, the empirical basis of this necessitation aspect of causation is captured by the principle 'same cause, same effect'. If we have two situations identical with respect to the precise cause and relevant environment, they both have the effect occurring. It embodies the pragmatic upshot of causal theorizing, as it allows us to draw straightforward inferences from empirical observations to general rules or laws about causation that we can then use to achieve a desired effect by creating the appropriate cause in a suitable environment. Yet, the 'same cause, same effect' principle by itself is practically vacuous because it doesn't indicate the relevant factors for judging whether two given causal situations are similar enough to count as effectively the same cause.

It turns out fortunately that there exist structures in our fundamental physical theories that we can use to clarify these conditions: dynamical laws establishing nomological determination relations between microstates at different times. With these laws in hand we can interpret 'same exact cause' to mean 'same local microstate,' and 'same relevant environment' to mean 'same exact microstate'. In this sense, deterministic dynamical laws vindicate our thinking of the world as obeying the 'same cause, same effect' principle. Crucially, the microstructural relations allow us to determine how much difference in the effect exists between two causal situations that have only approximately the same cause.

The folk concept of causation applies not to microstates, but to macroscopic objects, events, and processes. To make the connection between these entities and the nomological determination relations, we group together some of the local physical facts that reasonably fall under some convenient description like ‘striking a match’ or ‘neuron firing’. In many circumstances, such an underdescription of the physics serves well as a proxy for the microscopic physics that in practice is epistemologically inaccessible and too complicated for making inferences. For example, we say ‘neuron *A*’s firing causes neuron *B*’s firing’ to describe physical situations that could be more accurately described as ‘the *A*-microstate nomologically determines the *B*-microstate’. The causal terminology omits reference to physics outside the neuron even though the micro-facts in one’s liver are needed to necessitate *B*’s firing just as much as the micro-facts in *A*. This omission is often excusable because usually *A*-microstates with minor differences in the liver nomologically determine microstates that also include *B*’s firing. That is, the vague language is good enough when the physics in the area around *A* and *B* is sufficiently insensitive to the kinds of physical facts one typically finds further away in the liver. But because sensitivity to the external physics is a matter of degree, no matter how you individuate these coarse-grained events, there are always going to be borderline cases where the causal terminology substantially misrepresents the microphysics. This is a key source of intractable difficulties for robust theories of causation, which try to defend much more of the folk theory of causation.

The justification for counting “neuron *A* causes neuron *B* to fire” as a good approximation of the real physics is that the physics of neuron *B* is more sensitive to the physics of neuron *A* than it is to the liver. In order to measure sensitivity objectively we need to compare various microscopic modifications to the initial state, which requires that we have some objective measure over the possible *A*-microstates and their nomological determinants. Fortunately, we have objective probability measures in the theory of statistical mechanics that can arguably quantify sensitivity by allowing us to compare counterfactual macrostates by quantifying what proportion of their microstates lead to what effects. We justify the focus on the *A*-microstate as the cause by noting that the fraction of alterations to the neuron-*A* part of the microstate leading to *B*’s not firing greatly exceeds the fraction of alterations to the liver part leading to *B*’s not firing.

Once we have statistical mechanical probabilities in our theory of causation, it is easy to add fundamental dynamical probabilities, the kind that exist in stochastic theories where nature makes random jumps according to probabilistic rules. When we have this kind of indeterminism in our theory, typically we no longer have a ‘same cause, same effect’ principle, but instead a ‘same cause, same probability of effect’ principle, which works well enough in many circumstances.

Incorporating statistical mechanical probability is also good because it allows us, even in a deterministic environment, to associate causal processes with probability-raising processes, widely-recognized as a decent first-order approximation to the folk theory of causation. Accounts of causation as probability-raising are also known to have serious problems like pre-emption and fizzling (Schaffer 2003 [10]), but such problems, even if insoluble, need not count decisively against the theory being partially developed here because it already admits that the underlying physical principles may not perfectly capture folk intuitions. What will eventually be needed is an explanation of why other grounding principles of causation interfere in cases like pre-emption and fizzling to override the rule of thumb that causation involves probability-raising.

We are now in a position to explain why it is often fruitful to conceive of causation as transitive. Whenever the *A*-microstate nomologically determines the *B*-microstate and the *B*-microstate nomologically determines the *C*-microstate, it follows that the *A*-microstate nomologically determines the *C*-microstate. That is, we have transitivity at the level of microscopic determination. Putative counterexamples to causation involve the coarse-graining of events that prevents the straightforward application of transitivity. For example, Schaffer (2003 [10]) (originally from Hall 2000 [4]) describes *A*, a boulder rolling down the hill towards a hiker's head, which causes *B*, the hiker to duck, which in turn causes *C*, the hiker's survival. If causation is transitive, we are forced to accept the apparently counterintuitive claim that the boulder rolling towards the hiker caused him to survive. It's easy to see why it is counterintuitive at the level of coarse-grained description: *A* makes *C* less likely in the sense that there is a bigger proportion of boulder-rolling-microstates that determine the hiker's death than the proportion of boulder-stays-still-microstates that determine the hiker's death. This probability lowering comes despite the fact that *A* raises the probability of *B*, and *B* raises the probability of *C*. The coarse-grained description focuses our attention on the fact that there is a rolling boulder, which lowers the probability of survival and hence counts as at least some kind of reason for denying that *A* causes *C*. Nevertheless the particular boulder that rolled was the most significant part of a larger microstate that determined that the hiker survived, which gives us a reason to say *A* did cause *C*. The fact that such purported counterexamples to transitivity are seen not to be counterexamples when we describe the same facts at a more fine-grained level, strongly suggests that the justification for the causal transitivity ultimately lies in the nomological determination part of causation and is sometimes obscured under a more coarse-grained description of reality.

2. Causal Asymmetry

The other prominent aspect of causation is the causal asymmetry, the fact that causes temporally precede their effects. It is very tempting to conceive this asymmetry as somehow embedded in the local physics—that there is something in the striking of the match that made it burn but nothing in the burnt match that made it previously struck. However, a careful look at the underlying physics does not support the idea that the causal asymmetry is localized. Surprisingly, the apparent asymmetry between the striking and the burning is grounded instead by way of special boundary conditions of the early universe.

2.1. **Fundamental Physical Asymmetries.**

To seek the physical ground of the causal asymmetry, we need first to examine the kinds of temporal asymmetries existing in plausible fundamental physical theories. There are two possible types worth consideration here. The first is a temporal orientation, which locally defines one direction in time as dynamically different from the other. The most common version of this temporal orientation is implicit in the use of a stochastic dynamics. In a theory with stochastic dynamics, one has laws of nature specifying that chance processes sometimes occur and that any time a chance process occurs, the physical state on one temporal side of the chance process depends in a fundamentally probabilistic way on the physics of the other temporal side. The independent side is what we call 'the past' and the dependent side, 'the future'.

Indeterministic and stochastic dynamics of the kind that have been proposed as serious fundamental physical theories look superficially like they might give a plausible explanation of the causal asymmetry. Because a stochastic dynamics has rules for calculating the probability of future microstates from the current microstate, there is a determinate probability for what would have happened had a given cause not occurred. Any full microstate without the cause has some nomologically determined probability for the effect occurring. Thus, there will be facts of the matter about what local chunks of matter raise the probability of any given effect.

However, even though stochastic rules make the future chancy, all the kinds of stochastic laws that appear in realistic fundamental theories—like laws about quantum mechanical wave collapse—fail to make the past determinate. In fact, they don't constrain the past at all, and so restrict the past even less than they do the future. This has the consequence that they cannot, by themselves or together with other deterministic dynamical laws, justify an objective probability measure over the past states compatible with some hypothetical present state. The probabilistic rules that go only from present to future cannot be applied in going from present to past unless we have some independent grasp on the initial probability distribution, which is not given by the dynamics. So the temporal orientation determined by the stochastic dynamics, far from explicating causal asymmetry, makes the problem worse by making the past less fixed than the future.

However, even though the stochastic rules make the future chancy, they fail to make the past determinate. In fact, because they do not constrain the past at all, they restrict the past even less than the future. This has the consequence that they cannot, by themselves or together with any other dynamical laws, justify an objective probability measure over the past states compatible with some hypothetical present state. The probabilistic rules that go only from present to future cannot be applied in going from present to past unless we have some independent grasp on the initial probability distribution, which is not given by the dynamics. So the temporal orientation determined by the stochastic dynamics, far from explicating causal asymmetry, makes the problem worse by making the past less fixed than the future.

The other kind of physical asymmetry is an asymmetry in boundary conditions. For example, we know there is a smooth, bunched up distribution of matter and energy at the temporal end of our universe we call the past, and another clumpy, spread out distribution a good distance into the other temporal direction we call the future. A long history of investigation into the foundations of statistical mechanics indicates this difference cannot be explained in terms of the dynamics. If we take for granted all the gross features of the physical state in the early universe, the dynamics does tell us that matter at the other end of time will be spread out and clumped. However, if we take for granted the macroscopic features of the physical state at some future time, we cannot infer the existence of the smoothly concentrated matter of the early universe. This is true even though a deterministic dynamics entails that the exact future state determines every feature of the early universe. The asymmetry exists at the coarse level of description where we have less than a full microscopic state to use for inferences.

There are many kinds of physical asymmetries that cannot be explained by dynamical asymmetries but only by boundary conditions. For example, we have thermodynamic asymmetries in the dispersal of gases and the flow of heat. These thermodynamic asymmetries are grouped together theoretically in that they all can be summarized by the rule that the entropy of an isolated system virtually never decreases (as we go forward in time). The dynamics cannot explain thermodynamic asymmetries because from the point of view of the microphysics, the kinds of dynamical behavior one needs to repeatedly drive a wide variety of systems towards lower entropy are far too variegated. One can see this by way of standard examples in the literature on statistical mechanical explanations of entropy increase. Consider a gas inside an isolated tank, where we idealize the gas as molecules interacting only by elastic collision, and we take the gas at $t = 0$ to be uniformly distributed only in a small volume V of the tank. Because the gas is uniformly distributed, there will almost certainly be many gas molecules on the edge of V that happen to have their velocities pointed towards the empty space in the rest of the tank. These molecules will spread out over the first few seconds and after a short time the gas will be uniformly distributed throughout the tank and will stay there at equilibrium. Let S label the microstate of the gas and tank at $t = 1$ min. We know that there is a physically possible microstate S^* which is just S with all the particle velocities reversed. The classical dynamics makes S^* evolve in a way that is macroscopically the reverse of S, so that S^* will sit at equilibrium for almost a minute, and then the particles will hit each other in just the right 'improbable' combination to make the gas collapse to the small volume that S started in.

The question we are interested in is whether dynamical laws acting at a local level could explain the behavior of gases macroscopically like S^*. While one could cook up a dynamics that has certain chance moments where dispersed gases collapse, such dynamics would fail to reproduce the relevant behavior of S^*'s evolution. To see this, augment the example by having V be the inside of a small canister inside the tank that is opened to let the gas escape. For an embellishment also imagine that there are other canisters in the tank with small leaks making them useless for holding gas. Picturing the evolution of S^*, we have the uniformly distributed gas doing nothing for almost a minute and then collapsing into the one functioning container and being sealed in by an apparently spontaneously shutting lid. In this example, the dynamical law, even if it had spontaneous gas collapses, would not only have to collapse the gas but would have to collapse it into the one leak-free container. It would need to be responsive in a reliable way to facts that are very hard to describe in any way other than as macroscopic facts about will happen in the future. The particle motions would need to conspire to be in just the right microscopic condition to instantiate future states that are overwhelmingly unlikely but satisfy a compact macroscopic description. This kind of responsiveness, local dynamical laws do not have.

The correct explanation of thermodynamic asymmetries comes by way of an asymmetric boundary condition. Specifically, we posit low entropy in the past and no similar constraint in the future. This entropy constraint on the macroscopic state of the entire universe often goes by the name 'the past hypothesis' (2000 [1]). It follows from the past hypothesis in addition to the dynamics and standard probability measure we already have, that the universe will likely exhibit thermodynamic asymmetries of the kind we regularly see. It is exactly the structure of such boundary constraints that makes dynamical evolution of a macrostate towards the past exhibit the seemingly conspiratorial motion that we in fact see happen in reverse. And it is this feature that explains the seemingly local character of the causal asymmetry, as we will soon see.

2.2. **Counterfactual Asymmetry.** In order to apply causal terminology to coarse-grained events in a way that justifies our selection of some events as salient, we needed a structure that made possible a measure of the causal sensitivity of various chunks of physics. We found this measure in the theory of statistical mechanics. We can equivalently treat our comparison of various microstates as a comparison among counterfactual possibilities, the possible worlds that possess the microstates in question.

The recourse to counterfactuals in elucidating causation is familiar, and suggests the possibility that the causal asymmetry can be fully explained as a counterfactual asymmetry. For example, the asymmetry of causal dependence between the striking and burning of the match can be expressed as, "It's true that had the match not been struck, there would have been no flame, and it's false that had there been no flame, it would not have been struck." In making this idea precise, it's worth noting that our offhand judgments about counterfactuals do not universally justify such an asymmetry. In a significant number of cases we judge that if the effect had not happened, its cause wouldn't have happened either. For example, we get counterfactual dependence of the past on the present when we imagine the cause and effect as part of a fixed system. A ring falls into the sink, bounces around randomly and slips past the drain cover, causing it to land in the plumbing. It's reasonable that if the ring were not in the plumbing, it would be because it didn't slip past the drain cover. We also get backwards dependence in more theoretically-minded evaluations such as when we focus on the underlying deterministic microphysics. When an A-microstate nomologically determines a preceding B-microstate, it is reasonable to claim that B would have been true if A had obtained in the exact way specified by the A-microstate.

Because our evaluation of counterfactuals can often in ordinary circumstances fail to possess the asymmetry needed for causation, to explain the causal asymmetry by the counterfactual asymmetry we need either a theoretical refinement of counterfactual reasoning or else a way to understand the causal asymmetry as only loosely tied to the counterfactual asymmetry. I'll first consider two different justifications for seeking a causation-friendly refinement of our ordinary counterfactual reasoning and dismiss them as insufficient. Then, I'll go on in section 2.3 to elaborate a theory that treats the counterfactual asymmetry as only a rough approximation, allowing the so-called backtracking reasoning associated with counterfactual dependence of the past on the present. Nevertheless, even without a strict counterfactual asymmetry, we will have enough of a counterfactual asymmetry to justify the causal asymmetry for all practical purposes.

The most famous attempt at the refinement strategy is David Lewis's "Counterfactual Dependence and Time's Arrow" (1979 [7]). Lewis recognizes that some contexts allow the counterfactual dependence of the past on the present, but distinguishes what he calls the 'standard resolution'. The standard resolution of counterfactuals is stipulated to disallow the undesired counterfactual dependence of the past on the present. To justify the appeal to a standard resolution, Lewis presents a theory of counterfactual evaluation that tries to clarify the criteria we effectively have in mind (or should have in mind) when we think about counterfactuals. His theory tells us which respects of similarity to use in comparing possible worlds and is designed so that the most similar worlds with the antecedent A true (assuming A is actually false) will turn out to be worlds with the exact same microscopic past up until a short time before the events mentioned in A happen and afterwards a different but lawful future. A legitimate review of Lewis's theory deserves more attention than I can give here, so I will set it aside with the suggestion that his theory, if it were patched to avoid all the known counterexamples (e.g., Edgington 1995 [3] for a review and Schaffer 2004 [11] for discussion of some more recent examples), would be too baroque a system to count as a justification of the standard resolution. Furthermore, Lewis's theory comes with no explanation of why our local physical environment satisfies the 'overdetermination asymmetry' he posits as the explanation of counterfactual asymmetry. In section 2.3, I discuss why the past hypothesis and dynamical laws often establishes something like Lewis's overdetermination asymmetry, and with this physical explanation in hand, Lewis's recourse to miracles and the associated priority rankings is made superfluous.

A second potential justification for concerning oneself only with the standard resolution is that decision-making situations seem to require it—that when deciding among alternatives, one shouldn't use backtracking reasoning. The hypothesis here is that the counterfactual asymmetry needs only to be justified for the kind of situations where we humans are able to influence the world by way of decisions. Then, we project this asymmetrical aspect of our agency to other objects in nature, making causation seem asymmetrical when it is really only decisions and our perspective that are asymmetric.

In one example (Downing 1959 [2, 7]), Jack and Jim have a fight and the next day Jim considers asking a favor of Jack, but decides not to. There are two reasonable ways to understand what would have happened had Jim had asked the favor. On one reading, Jack would refuse because he is still angry. On another reading, we fix on Jim's pride as a salient characteristic to keep constant and infer that if he were to ask Jack for a favor, it would have been because there was no fight the day before. Then we reverse direction and infer that if there hadn't been a fight, Jack wouldn't be angry and so would likely grant the favor.

The reason decisions appear counterfactually special is that it is pragmatically irrational to use backtracking reasoning when making decisions. It is exceedingly unwise for Jim, who remembers the fight quite well, to use the backtracking reasoning that leaves subject to his volition whether yesterday's fight occurred. He shouldn't think (indicatively) that if he asks Jack, Jack will grant the favor because there wasn't a fight. Due to the prevalence of situations like these, it is arguably an empirical fact that people who use such backtracking reasoning will typically be less successful in achieving their aims than those who use the standard resolution, *ceteris paribus*.

Regardless of the plausibility of justifying the standard resolution of counterfactuals in this way, the idea that we project the counterfactual asymmetry onto the world is undermined for two reasons. First, counterfactuals involving decisions can sometimes be naturally interpreted in backtracking-friendly ways, like decisions that are psychologically difficult or decisions to accomplish tasks that only make sense in other contexts. For example, if I had decided yesterday to convert to Asatru, a religion I know nothing about, that would have been in part because I had earlier learned something about Asatru. If I had decided yesterday to descend Denali, it likely would have been that I had previously ascended. The reason decision-counterfactuals don't usually involve significant backtracking is that we usually imagine decisions where a person is already restricting the options to those that are possible given what he or she knows.

Second, limiting the scope of our explanation to decisions misses the point. The goal is to explain why *the world* is friendly to a folk physics that invokes causation. It is not enough to explain why the counterfactual or causal asymmetry applies merely to decisions. We need the explanation to apply to the wider environment, so that we can explain why it does us no good to try to cause events by way of processes that take place after the fact. The asymmetry demanding explanation is the temporal asymmetry in the kinds of causal chains we can hope to create.

2.3. **Counterfactual Asymmetry via Boundary Condition Asymmetry.** Consider the following explanation of a counterfactual asymmetry that permits backtracking and counterfactual dependence of the past on the present while still giving us a kind of causal asymmetry. The basic idea is to evaluate counterfactuals dealing with physical events by way of the objective chance that the consequent would obtain among the relevant possible worlds where the antecedent obtains. The counterfactual conditionals of primary concern are ones where the antecedent A describes facts that can plausibly be thought of as some coarse-grained events localized in some spatial region R at time t. For these conditionals, take the microstate of the actual world at t, consider all the localized microscopic modifications needed to make A obtain, and then let the objective probability distribution and the dynamics tell us how likely it is for the consequent C to obtain. One can think of this objective conditional probability $prob(C/A)$ as a measure of the degree of assertibility (or acceptability) we ought to have in the conditional if we knew all the actual facts, or one can use the standard semantics for counterfactuals and claim "$A \mathrel{\Box\!\!\rightarrow}$ the objective chance of C is $prob(C/A)$". This evaluation procedure extends to more vague counterfactuals like "If people ate more vegetables, there would be fewer cases of diabetes," by thinking of them as generalizations about more specific counterfactuals, i.e., what would happen if particular people ate particular vegetables at particular times.

In this theory, what it means for the past to be fixed is for it to turn out (for almost all actual coarse-grained past facts, F, and almost all localized antecedents A) that "$A \;\Box\!\!\rightarrow$ the objective chance of F is very high." The explanation for why the past is fixed is that when we counterfactually suppose A, we include two important (but defeasible) background assumptions in addition to the dynamical laws. We assume that the past hypothesis PH is true and that microscopic facts M outside of R are held fixed (merely because we are restricting consideration to localized counterfactual alterations). For ordinary antecedents A, making explicit the tacit hypotheses means $prob(C/A)$ should be understood as $prob(C/A\&PH\&M)$. The effect of having the constraints M and PH is plausibly that lots of macroscopic past facts are highly probable given these two constraints. It is important to realize that whether this is true for any particular antecedent depends on the actual dynamics and the antecedent itself, and unfortunately reality is far too complicated for us to test whether particular counterfactual assumptions really do hold the past fixed. So we can proceed only by making intelligent guesses about how the dynamics will work under the given constraints using the most reasonable rules of thumb available.

A quick sketch of an example should convey how PH and M team up to fix the past. For major historical facts, like Napoleon's rule of France, we have existing macroscopic evidence spread over a large spatial region. If we try to imagine what the most likely way for a world starting out like ours to evolve into a world with all this widespread evidence, it is reasonable to guess it would be largely through worlds where Napoleon did really rule France and not through worlds with accidental accretions of misleading evidence. Minor historical facts lack macroscopic traces, but if we look at the present detailed microstate, it is plausible that we have lots of microscopic bits of evidence like images streaming away from Earth, etc. These microscopic traces could plausibly conjoin to make highly likely the historical events they seem to jointly imply. The presence of micro-traces also adds credibility to inferences based on the existence of macroscopic traces. While macroscopic evidence is sometimes misleading by way of hoaxes, cover-ups, accidents, etc., it is difficult to hide all the macro-traces without leaving at least microscopic traces of the cover up. Insofar as we have at least microscopic traces (at t) of a previous macroscopic fact F, holding fixed the macrostate of the early universe and the present microstate outside R makes F very likely to occur in the worlds where the antecedent is true.

An important caveat is that there is no fact of the matter in general about how large R should be. To some extent our ordinary reasoning about counterfactuals involves inferring from facts inside R at t to previous times and then backtracking to t in a way that might require us to adjust the microstate outside R. For example, we might start out thinking that in order to make Jim ask the favor, we need only adjust the physics of Jim's head and leave untouched the microphysics elsewhere. But if we keep fixed Jim's pride, then there would likely have been no fight yesterday, and if there had been no fight yesterday, there would be microtraces of this fact in the present, so we are postulating microscopic changes outside R after all. This shows that at best the evaluation of ordinary counterfactuals involves delicately balancing the consequences of what we have in mind when we consider antecedents that can backtrack. The details of the full theory are covered elsewhere (Kutach 2001, 2002 [5],[6]).

The picture this theory is supposed to vindicate is that the counterfactual A-worlds are mostly worlds that start off microscopically very nearly identical to the actual world, with the motion of every particle in the counterfactual universes so nearly like the actual microstate that there is no noticeable difference until some reasonably short time before t. During that time, the microscopic differences very, very slowly build until some (usually brief) time before t, the differences become big enough to set off dynamical changes that quickly become macroscopic, eventuating in A being true. The counterfactual differences come about through normal lawful evolution of physics, and do in ways weighted probabilistically by the physics, so that there is no miraculous funny business. This picture supports our intuition that there is some limited counterfactual dependence of the past on the present, dependence insofar as the past had to be different in order for the present to arise from the past by way of the actual dynamical laws. Yet, the two constraints make counterfactual differences in the past almost entirely microscopic, so that virtually all the macroscopic truths about the past are effectively held fixed under localized counterfactual manipulations to the present. The probabilistic feature also supports our intuitions that usually forward-directed counterfactual processes, like actual processes, can be understood through the usual procedures of conditionalizing on known facts. For example, if we have a device with a chancy mechanism that 15 percent of the time rings a bell via causal process P, 45 percent of the time rings the same bell via a different causal process Q, and 40 percent of the time doesn't ring the bell, then we can conclude the following: If the bell doesn't actually ring, then the chance that P would have occurred if the bell had rung is 25 percent.

The past hypothesis plays a critical role because if we don't keep it as a constraint, the most likely worlds making A true and the microstate outside R fixed would be worlds with an anti-thermodynamic past. One can see this from the example 2.1, where S^* is evolving towards a condensed volume V. Due to the dynamical fragility of the evolution towards low entropy, it only takes a small deviation in the motion of a single particle to disrupt an evolution towards low entropy into an evolution towards high entropy. In the real world, there are lots of ordinary forces that can spread the effects of motion, so any localized disturbance will quickly spread to other parts of the universe. Thus, without the past hypothesis, the past would be far more sensitive to the present than the future would be.

This account of counterfactuals supports something like David Albert's (2000 [1]) explanation of the asymmetry of knowledge and causation. Although a long tradition exists where thermodynamic asymmetries are explained by the past hypothesis, attempts have failed to explain other asymmetries—like the fact that our knowledge of the past is more precise or more abundant than our knowledge of the future—by connecting them to entropy increase. Albert's key insight is that epistemological and causal asymmetries are not explained by the past hypothesis in virtue of the past hypothesis making highly likely the entropy increase of mental structures or causal sequences but instead are explained directly in terms of the past hypothesis itself, regardless of whether the entropy of mental structures or causal sequences is even definable.

In giving an account of causal asymmetry, Albert first defends an epistemological claim, that what we know about the world (assuming a classical statistical mechanical worldview) can in principle come by way of a certain inference procedure where we derive probabilities of facts from the conjunction of the laws, the past hypothesis, the objective probability distribution, and the current 'directly surveyable condition of the world', which he characterizes as the world's current macrostate plus possibly a few microscopic features one might be able to introspect (p. 96). The resulting epistemological asymmetry is connected to the counterfactual and causal asymmetry because when we consider alternative possibilities, we are usually concerned with worlds like our own, meaning worlds where this inference procedure is effective. Thus, we can use the inference procedure to make inferences about what counterfactually would have happened under certain postulated circumstances. Albert then argues that this justifies our believing that among all our causal handles, i.e., things over which we have immediate control, there are a 'far wider variety' that affect the future.

In my specification of $prob(C/A\&PH\&M)$, the past hypothesis, laws, and probability distribution play essentially the same role as in Albert's argument, although M is characterized in a way that fixes the past more. The two most important differences in the accounts are that my truth (or assertibility) conditions for counterfactuals (1) are explicitly independent of any epistemological claims, and (2) apply to counterfactuals beyond those involving volition and decisions. Regardless of differences, the following theory of causation can be interpreted as one way of making precise the sense in which our causal handles on the world are temporally asymmetric.

3. Causation

An event c is a *contributing cause* of event e if and only if c is an essential part of some microstate that nomologically determines an objective probability for e. For c to be an essential part means that the microstate doesn't determine the probability e if we exclude the fact that c occurred. From there, we can go on to make finer distinctions designed to determine how important any cause is compared to other contributing causes of e. Our folk notion of causation at least roughly tracks the notion of important cause.

One feature that usually serves as a good sign of c being an important contributing cause to e is that replacing c with some alternate physical state leads to E being significantly less likely to obtain, where E is some coarse-grained description of e. To make this more precise, consider all the possible microstates constructed by taking the actual microstate S at the time t when c occurs and modifying it by replacing the spatial region occupied by c with some physics that involves C not occurring, where C is a coarse-grained description of c. There are, in general, many different ways to make C not occur, depending partially on how narrowly C is characterized and partially on what other background assumptions we make about what kinds of non-occurrence are contextually relevant. Supposing there is some reasonable space of such microstates, let $C^\dagger$ be the proposition corresponding to the possible worlds that are nomologically compatible with these microstates. $C^\dagger$ captures the relevant sense in which c does not occur. A good (but defeasible and not fully objective) measure of whether c is an important cause of e is whether the probability of E is boosted more by C than by other things. We can measure the degree of sensitivity of E on C or the degree to which c is a probability-raiser for e, by way of $prob(E/C^\dagger)$, the objective probability that E occurs among the relevant worlds where C doesn't occur. If $prob(E/C^\dagger)$ is significantly lower than $prob(E/C)$ or is relatively low compared to other contributing causes, we can typically say that c is an *important cause* of e.

An example of simple causation is when Billy throws a rock at a bottle, breaking it one second later. Suppose that the event c_1 of Billy throwing the rock occurs at time t, as part of a microstate S that extends across a sphere of radius one light-second. Assuming some dynamics compatible with relativistic locality, the contributing causes are all the various chunks of S, including innocuous events like c_2, an ant crawling on the top of the Taj Mahal. Billy's throw is underwritten as a relatively important cause because the bottle's breaking, E, depends more on c_1 than on other contributing causes like c_2. The dependence of E on c_1 is the objective probability that the bottle will break in worlds that are just like actuality except with Billy not throwing the bottle. This is low by the presumption that nothing else in the environment makes the probability of E high, i. e., that there are no backups. E is not dependent on c_2 because virtually any reasonable way of making the ant absent from S will have negligible dynamical consequences for the bottle.

This defeasible marker for causal importance can also be expressed using counterfactual conditionals. We can say event e under the description E counterfactually depends on c if and only if had c not occurred, E would likely not have obtained. This way of expressing the dependence relies on a special interpretation of the counterfactual: that we interpret c's not having occurred in terms of an objective probability distribution over the $C^\dagger$ worlds, resolving the vagueness of "had c not occurred..." in a way that excludes backtracking reasoning.

Another important physical feature that plays a role in determining causal importance is the presence of intermediate determining microstates. In all important fundamental theories that physics has so far uncovered, when one state A nomologically determines a later state B, it does so while also determining any temporally intermediate states to be such that they also determine B, except possibly for esoteric cases that are irrelevant to ordinary causation. In common cases of causation, not only is $prob(E/C^\dagger)$ significantly lower than $prob(E/C)$, but in all the intervening states, there are events that are related to previous and later states by this same probabilistic connection in a chain or continuum. In cases where an intermediate determining state contains no events that connect to previous and later causes with this probabilistic relation, we think of this as a case of a broken causal link and thus have a good reason to reckon c as unimportant. Important causes, the intuition goes, deliver their importance through a causal process. (See also, Schaffer 2001 [9])

These considerations guiding our evaluation of causal importance are imperfect and far from exhaustive. But since my aim here is not to give a full account of causation, I will just suggest that evaluating whether c causes e by measuring *importance* offers some benefits over more robust analyses. First, because conflicting kinds of importance might bear on causal judgments, causation could be treated as sometimes involving interest-relative features. This may prove useful in acquiring a better fit between the theory of causation and judgments of culpability. Second, different kinds of causal importance might have irreconcilable differences, and if we can identify the source of conflict, we might resign ourselves to the presence of seemingly persuasive counterexamples that we cannot accommodate. Third, by shifting much more of the pragmatics of causation away from the physics, we can allow greater discrepancies between our theory of causation as it is in the world and causation as it seems to us. This lets us violate ordinary convictions about causation in order to achieve peace with physics while nevertheless accounting for the usefulness of problematic folk convictions. An instructive example is the case of causal asymmetry.

4. Causal Asymmetry Revisited

The conditions for a contributing cause involve only facts about nomological determination and involve no temporal asymmetry that makes the past fixed. Yet we can find a causal asymmetry in the criterion for important contributing causes. It arises from the use of the past hypothesis as a background condition implicit in $C^\dagger$. Assuming the arguments in section 2.3 are good, the past hypothesis will fix most macroscopic facts about the past under localized counterfactual suppositions about the present. Supposing that we have a fully (bidirectional) deterministic dynamics, my wiggling of my finger, c, constitutes a contributing cause (on the theory so far) of Napoleon's invasion of Russia, E, but because

$prob(E/C^{\dagger})$ is nearly one, c doesn't count as an important cause. Indeed, present contributing causes of the past are so often unimportant, it is convenient and almost always permissible to ignore this kind of backward causation and just imagine causation to have an asymmetry where only earlier events can be contributing causes to later events.

However, even though the past hypothesis by and large fixes the past, there exist cases where the past counterfactually depends on the present, and these cases threaten to count as instances of backward causation. Nevertheless, it turns out that even where the past depends on the present, we are unable to exploit this dependence to accomplish anything practical. Hence, we are for practical purposes justified in assuming a uniform causal asymmetry and explaining away the limited counterfactual dependence of the past on the present in ways that are causally innocuous. I consider three cases below.

4.1. **Common Causes.** The treatment of counterfactuals in section 2.3 permitted backtracking reasoning in the evaluation of what would have happened had c not occurred. This seems to license a mutual counterfactual dependence of two events on each other, and thus mutual causation, in cases where we would ordinarily say that the two events are merely effects of a common cause. For example, let v be the infection of a person with some specific virus, r be the rash that appears one day later made highly likely by v (in the sense that $prob(R/V^{\dagger})$ is low), and let f be the fever that appears two days later made highly likely by v in the same way. Imagine there are no other likely mechanisms by which R or F can occur, e.g. the patient is isolated, the rash has the signature color of this virus, etc. In this case, it is perfectly reasonable to claim counterfactual dependence of F on R, that if the patient hadn't gotten the rash, she wouldn't have gotten the fever on the grounds that there is no other plausible way to get the rash other than by having the virus. Nevertheless, counterfactual dependence is not a sufficient condition for one event being an important cause of the other, even when conjoined with nomological determination.

The importance of r as a cause of f is measured by $prob(F/R^{\dagger})$, which is the likelihood that the fever occurs, given that the patient is in a state just like actuality but with no rash. Hypothetically removing merely the rash from the patient at some time t does not necessarily involve removing the virus as well. The microstates that determine $R^{\dagger}$ are all those molecular configurations where the blood vessels near the skin are smaller, but factors in the blood are left as is. Under the presumptions in the example, the virus would still be there in the body to make F likely, and so $prob(F/R^{\dagger})$ would be high, and so the rash doesn't count as an important cause of the fever. If one argues that for some reason the virus needs to count as part of the rash, then the low value for $prob(F/R^{\dagger})$ will signal that the rash-virus combination is an important cause of the fever, but this is no surprise since the rash-virus *is* an important cause of the fever in virtue of its viral part.

4.2. **Faking Traces.** Because traces in the future make highly likely the events of which they are traces, one might think we can affect the world by way of affecting what traces exist, and thereby backwardly cause an effect by way of creating the traces that occur afterwards. Indeed we *can* do this, but it always turns out to be merely a disguised case of what we have heretofore understood as ordinary forward causation.

We can safely suppose an absurdly generous upper limit on our power—that we have no more control over the world right now than we would have if we could freely instantiate any microstate of our choosing compatible with the dynamical laws and the past hypothesis. Suppose I try to use this power to instantiate a state right now that will cause traces ten seconds from now (in the future) of my having thrown a stone into a pond five seconds from now, and thereby make it likely that the stone entered the pond. The existence of traces, the concentric outward traveling ripples on a smooth pond surface together with a stone bearing my fingerprint at the pond's bottom, etc. do indeed make it likely that the stone entered the pond, but we can distinguish two distinct classes of dynamical evolution that start with my choice of microstate and eventuate in the traces. The first class includes all those evolutions that include the stone not entering the pond, i.e., worlds where the traces are misleading. Instantiating a microstate that causes *misleading* traces does nothing to make the stone enter the pond, because the stone does not in fact enter the pond. The second class includes all those evolutions that involve the stone entering the pond. These are worlds where the dynamical evolution flows from my initial choice of microstate through the stone entering the pond and later to the ripples, etc. All these worlds are just worlds we interpret as having the ordinary causal order, where I have caused the traces of the stone entering the pond by way of throwing the stone. Since I am unable to affect events usefully in a backwards way even when granted the extraordinary power to instantiate the microstate of my choice, I am equally unable in my more humble actual circumstances. And since the world cannot be controlled by way of such backwards causation, we can safely ignore it.

4.3. **Single Trace Causation.** The overall counterfactual fixity of the past arises through the joint effect of the past hypothesis and the existence of traces in the present, but the *recent* local past is still usually counterfactually dependent on the present. This results from the normal, lawful evolution of the world into the counterfactually postulated state. For example, there are no apples nearby, but if I were to see an apple right now in front of me, it would be that the apple existed in front of me a millisecond ago. This comports with the principle that current situations evolve naturally out of previous circumstances, with nothing miraculously teleporting into view or springing into existence ab nihil. (I am not absolutely ruling out such phenomena but presuming that any adequate theory of counterfactual evaluation will make it such that in ordinary circumstances they are highly unlikely.) The question is whether this short-range counterfactual dependence should be understood as a kind of backward causation, and the answer is that technically, according to the rule for measuring which contributing causes count as important, it is a case of backward causation. Nevertheless, it cannot be exploited to do anything practical, and so we can safely ignore it.

Suppose my control over the world comes by way of some limited control over what is going on in my head, what you might call an act of will, w, falling under a coarse-grained description W. The choices I could have made but didn't are $W^{\dagger}$, the worlds possessing microstates where (1) I choose something other than W, (2) the facts outside my head are just as they actually are, and (3) the past hypothesis holds. By the criterion for important contributing causes, the kind of past event over which I can have control will satisfy some event type E, such that $prob(E/W^{\dagger})$ is low. Circumstances where $prob(E/W^{\dagger})$ is low are situations where there are insufficient traces in the present to force the high probability of E. For simplicity, we can imagine this situation as one where some event e has a single trace W in my mind with all other potential traces of e being shielded out by a physical barrier that blocks traces.

Assuming present traces really do counterfactually fix most macroscopic facts about the past, the macro-facts outside the shielded region, including macro-facts prior to the shielding's existence, cannot imply or make probable e's occurrence. Whether e occurs must as a consequence depend crucially on the microscopic details of the prior physical situation, just as macroscopic facts about who wins the lottery depend critically on microscopic facts about particular ping pong balls. This implies that the only kind of *macroscopic* counterfactual connection that e has to other parts of the universe come by way of w. There are different ways to interpret this situation. On one way of looking at it, w does backwardly cause e to occur, but since e has no causal connection to anything else at the macroscopic level, we are unable to gain knowledge of its effects or use the causal connection to accomplish aims other than causing events inside the shielded region. Hence, knowledge of this kind of causal connection is useless. On another way of looking at it, e dynamically arises from the chaotic microstructure of the universe, and any traces in my head that e depends on can be understood as merely a reliable detection of e. A reliable detector's state, after all, depends counterfactually on the state of the detected phenomena, but this in itself does not imply that the detector *causes* the phenomena to be what it is. Given the lack of practical utility for conceiving of the counterfactual dependence as backward causation, it is more convenient to think of the causation as unidirectional and interpret the dependence as a mundane instance of detecting e's occurrence.

5. Conclusion

The objective structure underlying causation, i.e., what the physics appears to tell us exists, is a local nomological determination relation among physical states at different times, possibly with additional stochastic relations, neither of which contains the kind of time asymmetry able to explain the asymmetries that occur in ordinary cases of causation among salient (coarse-grained) events. There is another feature of the physics that can explain the kinds of asymmetric phenomena in causation, features about the universe's boundary conditions. The boundary conditions do not support a counterfactual asymmetry at the microscopic level, but at the level of coarse-grained events, there is an important (to us) difference between the way past events and future events typically counterfactually depend on the present. Given a very strong constraint on the physics at the end of time we call the past, coarse-grained events in the past tend to depend less on variations of the present state than do similar future events. More precisely, past events are usually resilient under counterfactual changes to the present that involve only small, localized modifications, or changes that seem like they can easily arise from small modifications in their recent past. Furthermore, the boundary condition implies that the evolution of the physical state towards the past is highly conspiratorial. As a consequence, the kinds of actions we are capable of performing either have no effect on the past or they have an effect on the past that is unpredictable and uncontrollable. Thus, whatever backwards causation does exist is of no practical value and is safely ignored by our ordinary concept of causation. Because the only kind of manipulable causation is forward causation, we are pragmatically justified in projecting the rough counterfactual asymmetry into the world, treating it as a universal asymmetry localized in material processes. Causation is ultimately our amalgamation of the nomic determination structure in the fundamental physics and our misconceiving the manipulability asymmetry as somehow built into the local physics.

References

[1] Albert, D. *Time and Chance*. Cambridge: Harvard University Press, 2000. 6, 10

[2] Downing, P.B. 'Subjunctive Conditionals, Time Order, and Causation', *Proceedings of the Aristotelian Society* **59** 125-140, 1959. 7

[3] Edgington, D. 'On Conditionals', *Mind* **104** (414), 235–329, 1995. 7

[4] Hall, N. 'Causation and the Price of Transitivity', *Journal of Philosophy* **97**, 198–222, 2000. 4

[5] Kutach, D. *Entropy and Counterfactual Asymmetry*, PhD. Dissertation, Rutgers, 2001. 9

[6] Kutach, D. 'The Entropy Theory of Counterfactuals', *Philosophy of Science* **69** (1), 82–104, 2002. 9

[7] Lewis, D. 'Counterfactual Dependence and Time's Arrow', *Noûs* **13** (1979), 455–76, reprinted in *Philosophical Papers, Volume 2*, Oxford: Oxford University Press, 1986. 7

[8] Russell, B. 'On The Notion of Cause', *Proceedings of the Aristotelian Society* **13** 1–26, 1913. 1

[9] Schaffer, J. 'Causation, influence, and effluence', *Analysis* **61** 11-19, 2001. 12

[10] Schaffer, J. 'The Metaphysics of Causation', *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), `http://plato.stanford.edu/archives/spr2003/entries/causation-metaphysics/`, Spring 2003. 3, 4

[11] Schaffer, J. 'Counterfactuals, causal independence and conceptual circularity', *Analysis* **64** (284), 299-308, 2004. 7

Box 1918, Department of Philosophy, Brown University, Providence, RI 02912
E-mail address: `Douglas.Kutach@brown.edu`
URL: `http://www.brown.edu/Departments/Philosophy/Douglas_Kutach/index.html`