CAUSATION AND OTHER ASYMMETRIES IN TIME

Christian Loew

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy.

Chapel Hill
2013

Approved by:

L.A. Paul

Robert M. Adams

John T. Roberts

Marc Lange

Matthew Kotzen

ABSTRACT

CHRISTIAN LOEW: Causation and other asymmetries in time
(Under the direction of Laurie Paul)

We tend to think that the past brings about, produces, or shapes the future but not vice versa. Yet, most candidates for the fundamental physical laws are time-symmetric: these laws determine the evolution of the world in the forward direction, but they equally determine its evolution in the backward direction. I argue that, in light of this lawful time-symmetry, causation itself is bi-directional, that is, causation runs forwards but it also runs backwards. This view might sound absurd, but it follows from taking fundamental physics seriously. I argue that causation is law-governed, and so the time-symmetry of the laws grounds causation in both temporal directions. Moreover, my bi-directional view of causation is compatible with our experience. In fact, it provides a deeper understanding than previously had of why we can control the future but not the past and why scientific explanations are time-asymmetric.

ACKNOWLEDGMENTS

Thanks go first and foremost to my committee. Bob Adams has encouraged me to think deeper about the issues in this dissertation than I would have otherwise. The depth and breadth of his thinking as well as his passion for philosophy and his kindness will continue to inspire me. John Roberts kept providing me with detailed and insightful comments as well as vital encouragement. Marc Lange and Matt Kotzen have been extremely helpful, especially at later stages of the project. Above all, I like to thank my adviser, Laurie Paul. Laurie was the best adviser I could have wished for, giving me the space to develop my own ideas and helping me do so in every possible way. She will continue to be one of my role models as a philosopher and as a person.

I have been fortunate to be part of several wonderful philosophical communities. I like to thank the philosophy departments at the University of North Carolina and the University of Arizona, and the people who make them such fantastic places for doing philosophy. At the University of Arizona, Richard Healey was like an additional adviser for me. I like to thank him for the excitement and patience with which he taught me the foundations of physics. I have learned much about the art of thinking things through from observing him do so. I also like to thank the departments at MIT and ANU for being wonderfully supportive and welcoming during my visits. Special thanks go to Mike Bertrand, Finnur Dellsén, Luke Elson, Geoff Sayre-McCord, Jonathan Schaffer, Al Hájek, Ned Hall, Caspar Hare, Thomas Hofweber, Bill Lycan, Doug MacLean, Cole Mitchell, Ram Neta, Brad Skow, Susan Wolf, and Stephen Yablo.

Many great friends have supported me while writing. Special thanks go to Kristen Bell for believing in me; to Anna Troupe for support and encouragement; to Laura Britton for listening and cheering me up; to Jen Kling for much good advice; to Craig Warmke for his friendship and help in philosophical and other matters; and to David Glick for being a great friend and a great philosopher.

TABLE OF CONTENTS

LIST OF FIGURES

CAUSATION AND ITS PLACE IN THE PHYSICAL WORLD

Introduction

1 Motivation and overview

The dissertation is about causation, its temporal direction, and its relationship to control and explanation. I defend the view that causation is bi-directional, meaning causation runs both forwards and backwards. This view might sound absurd and far-fetched, but I shall argue that it follows from taking fundamental physics seriously and that it is also compatible with our ordinary experience. Moreover, I will show that my view leads to a deeper understanding than previously had of why we can control the future but not the past and why scientific explanations are time-asymmetric.

My view revises our ordinary understanding of the temporal direction of causation. I will argue that causation goes in both temporal directions but that it nonetheless is time-asymmetric because it has a different character in the forward than in the backward direction. This difference in character grounds the practical asymmetries associated with causation. In particular, I will argue that causation can come apart from control and explanation, and that forward causation supports our practices of control and explanation but backward causation does not.

There are three motivations for pursuing this revisionary account of the temporal direction of causation. First, my project is supported by a naturalistic-reductive approach to

metaphysics. According to this approach, rather than imposing our ordinary experience and intuitions onto a theory of causation, we develop our theory of causation in accordance with the structure and features given to us by fundamental physics, and only add features (like a privileged temporal direction) if there is some outstanding need to do so. I shall argue that causation is governed by the fundamental physical laws and that my bi-directional view of causation best fits the structure of these laws while, at the same time, it is compatible with our ordinary experience.

Second, my view resolves the puzzle of how causation fits into the physical world. Many philosophers have noted that our ordinary concept of causation fits poorly with how fundamental physics describes the world.[1] The most striking mismatch concerns temporal directionality. According to our ordinary view of causation the past determines the future in a deeper or more important sense than the future determines the past. But in accordance with the fundamental physical laws, the future determines the past in the exact same sense in which the past determines the future.[2] Thus there is a puzzle about how fundamental physics leaves room for forward-directed causation.

The literature contains several proposals for how the temporal directionality of causation can be grounded in fundamental physics despite this apparent mismatch. But while there are important time-asymmetries in fundamental physics (in particular, in the boundary conditions), no theory has shown how these differences vindicate the strict intuitive time-asymmetry of causation (cf. Price 1996 and Weslake 2006). My theory turns this failure into a virtue. The reason why we find no strict time-asymmetry of causation in fundamental

[1] Cf. Earman (1976a), Field (2003), Lockwood (2005), Norton (2007), Russell (1913), Price (1996), van Fraassen (1993).

[2] Cf. Albert (2000, chapter 2), Carroll (2010, chapter 2), Field (2003, 436pp); Greene (2004, chapter 6); and Lockwood (2005, chapter 9).

physics is that the direction of causation is not strict but merely gradual. I will show that bi-directional causation fits naturally with fundamental physics and thus secures a place for causation in the physical world.

Third, my bi-directional view of causation brings into sharp focus certain issues about control and explanation. It is natural to explain the time-asymmetries of control and explanation in terms of the causal asymmetry. Ordinarily, we think that we can control the future but not the past and that earlier events explain later events but not *vice versa* because causes precede but do not succeed their effects. This explanation only works, however, if we can justify why causes are relevant to control and explanation while non-causes are not. For instance, if the laws are deterministic in both temporal directions, then earlier events are lawfully determined by later events. So why can we not control or explain earlier events in terms of these later events? What is it about causes that makes them exclusively privileged for control and explanation?[3] Without an answer to such questions the account is not particularly explanatory because it treats causation as a 'black box' without a further story of what features that causes have and non-causes lack make them relevant to these practices.

My theory provides an illuminating story of why we can control the future but not the past and why earlier events explain later events but not *vice versa*. The key point is that causation is not as closely associated with control and explanation as we sometimes assume. For example, my arm movement causes very specific movements of certain air molecules but I have no control over the exact nature of these movements. My theory isolates the features that qualify some causal relations for control or explanation and shows that causal relations in the backward direction lack these features. While someone could endorse this account of

[3] There might be non-causal explanations, but we still think that there is a special type of explanation in which only causes can be cited but not events that, for example, merely lawfully determine an outcome.

the time-asymmetry of explanation and control without accepting that causation is bi-directional, my view of causation helps us see these features more clearly.

The dissertation has three chapters. The first chapter ("Blunting the arrow of causation") argues that the view of causation that best fits with fundamental physics is one where causation is bi-directional. Ordinarily, we think that causation has a strict temporal arrow. We think that the past shapes, produce, or brings about the future, but not *vice versa*. But the fundamental physical laws equally determine the evolution of our universe in both temporal directions. I argue that causation is governed by the fundamental physical laws such that the nature of causation is determined by the structure of these laws. So it is reasonable to think that lawful evolution in both temporal directions also grounds causation in both temporal directions. I defend this bi-directional view of causation and show that it is compatible with the time-asymmetries of control and explanation.

The second chapter ("Why we cannot control the past") gives a deeper account of our inability to control the past that is compatible with my bi-directional theory of causation. If I want to spend my next vacation in Paris, there is a lot I can do. I can make a hotel reservation, book a flight, etc. But if I want to have spent my last vacation in Paris, there is nothing I can do about it now. In general, our limited control over the future contrasts with a complete lack of control over the past. But why can we not control the past? Intuitively, we cannot control the past because our decisions do not cause past outcomes. A careful understanding of what it means for an agent to control an outcome shows that even if our decisions *did* cause past outcomes, we still could not control the past. Control in the relevant sense is more than just causal influence but requires an unobvious sort of knowledge, and we

would lack this knowledge even if our decisions did cause past outcomes. My account thus provides a richer model of what control is and what it would take to control the past.

In the third chapter ("Causation, physics, and fit") I focus on explanation to give an account of why our ordinary notion of causation that we use in the special sciences is useful despite its poor fit with fundamental physics. Recent work on causal modeling has deepened our understanding of causation and explanation. There is, however, a puzzle of why these causal models are successful, in particular given their time-asymmetry and locality. These models explain outcomes by showing how they depend on a relatively small number of localized, earlier variables. Yet, fundamental physics allegedly describes the world in terms of lawful determination between very global states and does not distinguish between the way in which the past determines the future and the way in which the future determines the past. I argue that, despite this apparent mismatch, we can explain why causal models are successful from the structure of fundamental physics. In particular, the same physical features of the world that explain the success of our local causal models also explain why it is a good idea for us to build time-asymmetric causal models.

2 Causation and the physical world

The three chapters fit together into a general view about the place of causation in the physical world. In the reminder of this introduction, I want to outline and motivate this view. It is extremely natural to think of the world as causally evolving, and a central aspect of our ordinary concept of causation is its temporal directionality. We can distinguish two aspects of this directionality:

(i) Causal Direction. Causal relations are directed from cause to effect such that *c* causing *d* is different from *d* causing *c*.

(ii) Temporal Direction. Causes often precede their effects, but effects do not (or at least not typically) precede their causes.[4]

Causal Direction says that each token of the causal relation has a direction, but it says nothing about their temporal orientation. For example, if all cause-effect pairs occurred simultaneously, then each token of causation would still point from cause to effect but causation would have no temporal direction. Temporal Direction adds temporal orientation by saying that causal relations often point in the forward direction but never (or not typically) in the backward direction.

This ordinary view of causation, however, allegedly fits poorly with how fundamental physics describes the world. Many philosophers of physics have held that not only does fundamental physics not contain any relation that corresponds to our ordinary concept of causation, but it does not even leave room for such a relation. The classic articulation of this view is Russell (1913), who argues that there is a drastic mismatch between our ordinary notion of causation and how fundamental physics describes the world.

This mismatch is most striking for the temporal directionality of causation. On our ordinary conception of causation, which involves both Causal Direction and Temporal Direction, the past determines the future in a way that has no analog in the backward direction. In contrast, Russell argues that fundamental physics does not distinguish between how the past determines the future and how the future determines the past. Specifically, most

[4] This formulation is meant to leave open the possibility that our ordinary causal concept allows for simultaneous causation and for isolated instances of backward causation, as many philosophers think it does.

candidates for the fundamental physical laws are deterministic in both temporal directions. That is, a full specification of the universe at any one time, together with the laws, entails both a unique future *and* a unique past. For instance, for a billiard ball that is sufficiently isolated from its environment, its position and momentum at any one time together with the laws entails its position and momentum at both later *and* earlier times.

Moreover, the problem does not significantly change if the laws are probabilistic as long as these laws have the same probabilistic character in both temporal directions; that is, if a complete specification of the universe at any one time, together with the laws, entails a probability distribution over all earlier and later times.

More generally, the problem is that physical determination seems to be a matter of the physical laws, and the laws provide the same kind of determination (probabilistic or strict) in either temporal direction. This bi-directional determination by the fundamental physical laws is allegedly incompatible with our ordinary, forward-directed conception of causation for the following reason. The fundamental laws, *qua* being fundamental, tell us the complete and exceptionless story of how our universe evolves over time. It is hard to see what room there could be for a causal asymmetry where the past determines the future more fundamentally than *vice versa*. It seems that if there were an asymmetry of determination, then it should show up in the laws. But since the laws contain no such asymmetry, fundamental physics leaves no room for causation.

Russell takes this mismatch between the fundamental physical laws and our ordinary concept of causation to support skepticism about causation. He argues that our ordinary concept of causation is misleading and that causal relations are not part of the objective physical world. Russell thus emphatically declares that “the law of causality […], like much

that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." (Russell 1913, 1)

Getting rid of causation, however, seems neither feasible nor desirable. In an important paper, Cartwright argues that objective causal facts are indispensable for underwriting the objective distinction between effective and ineffective strategies (cf. Cartwright 1979). She points out that it is an objective fact that, for instance, quitting smoking is an effective strategy for avoiding lung cancer but that having one's teeth whitened is not. This objective distinction is grounded in causal facts, viz., smoking causes lung cancer, whereas having yellow teeth is merely correlated with lung cancer. Cartwright argues that we therefore need objective causal facts to ground the objective distinction between effective and ineffective strategies.

Causation plausibly underlies other practices besides effective strategies, such as prediction, explanation, and control. One could similarly defend causal facts based on these practices. However, Cartwright's approach of bringing in effective strategies is particularly compelling for three reasons. First, the distinction between effective and ineffective strategies appears completely objective. Although what ends we desire depends on our interests, what strategies are or are not effective toward a desired end is a matter of fact that holds regardless of our interests. Second, it is entirely unclear how else to ground the distinction if not in causal facts. Effective strategies require a non-accidental dependence between events that goes beyond mere correlation, and it is hard to see what that dependence could be if not causation. Moreover, we do not know how to pick out this dependence in a way that does not already presuppose knowledge of causal facts.[5]

[5] This kind of circularity is widely acknowledged in the literature on causal inference. See, for example, Woodward (2003). See also Cartwright (1979).

Third, effective strategies are closely related to the goals of the special sciences. Effective strategies require the same distinction that also underlies the law-like regularities discovered by the special sciences. For instance, the same fact that accounts for why taking a certain drug is an effective strategy toward recovery also accounts for why drug intake explains the recovery. So causal facts are not just related to our everyday practices but also to scientific explanation. Because of this centrality, "abandoning the concept of causation would cripple science." (Field 2003, 435)

Cartwright and Russell's respective insights are in tension and thus create a puzzle. On the one hand, Russell argues that fundamental physics has no place for causal facts. On the other hand, Cartwright argues that "causal laws cannot be done away with, for they are needed to ground the distinction between effective strategies and ineffective ones." (Cartwright 1979, 420) So something has got to give. In a recent survey article, Hartry Field assesses that "the problem of reconciling Cartwright's points about the need of causation in a theory of effective strategy with Russell's points about the limited role of causation in physics […] is probably the central problem in the metaphysics of causation." (Field 2003, 443)

The puzzle has received significant attention in the recent literature.[6] Responses fall into one of three camps. First, Pragmatists agree with Russell that there is a poor fit between our ordinary notion of causation and fundamental physics. But they argue that our causal concept can be central to science and everyday life even if causal relations are not part of the

[6] See the articles in a recent volume (cf. Price and Corry 2007).

objective physical world (cf. Price 1996, 2007; and van Fraassen 1993). Pragmatists thus deny that the distinction between effective and ineffective strategies is fully objective.[7]

Second, Primitivists take causal facts to be objective fundamental constituents of our world that are not reducible to more basic entities. Some Primitivists argue that a proper understanding of fundamental physics shows that causal facts are a part of it after all (cf. Frisch 2005). Other Primitivists admit that causal facts are not part of fundamental physics but hold that we can consistently add them and that we have reason from outside fundamental physics to do so.[8] These primitive causal facts can underwrite effective strategies. Primitivists, however, face the challenge of showing how exactly the existence of these causal facts is compatible with the limited role of causation in fundamental physics.

Third, Reductionists hold that Russell was partly right insofar as there are no causal facts in fundamental physics. However, Russell was wrong in thinking that there is no causation. Rather, causation reduces to fundamental, non-causal facts. In particular, Reductionists argue that the temporal directionality of causation can be grounded by bringing in statistical facts from the boundary conditions in addition to the fundamental physical laws.[9] The emerging relation is supposed to closely fit our intuitive notion of causation.

In the dissertation, I defend a solution to the puzzle that is reductionist but in a new way. I argue that objective causal facts are grounded in fundamental physics, but to secure fit between causation and fundamental physical facts we need to revise our ordinary conception

[7] Price (1996, 2007) is the most developed such account. Price argues that the distinction between effective and ineffective strategies is "objective from our perspective" as agents (cf. Price 2007, 286).

[8] Cf. Cartwright (1979) and Tooley (1987). Maudlin (2007) argues that there are grounds for forward-directed causation in fundamental physics but that even otherwise we would have reasons from outside physics to assume causal facts.

[9] See Dowe (2000), Field (2003), Papineau (1985), and Reichenbach (1956).

of the temporal direction of causation. The metaphysics of causation that arises from fundamental physics is one where causation is not sharply directed but one where causation goes in both temporal directions and the difference between the directions is merely a difference in degree.

The mismatch between the fundamental physical laws and our time-asymmetric concept of causation is so bewildering because causation and the fundamental laws are closely related. We used to think of all laws as causal laws. So fundamental physics already contains something very much like causation, except that it runs in both temporal directions, viz., lawful evolution. I will argue that the most natural understanding of causation in light of fundamental physics is therefore that the fundamental laws ground causation in both temporal directions. On the face of it, this bi-directional causation is at odds with the time-asymmetry of effective strategies, as well as numerous other asymmetries that are associated with causation. However, I will argue that these asymmetries, properly understood, are compatible with bi-directional causation.

3 Characterizing bi-directional causation

Bi-directional causation is the view that causation runs both forwards and backwards. We have seen that our ordinary view of the temporal direction of causation involves two distinct claims: Causal Direction and Temporal Direction. My bi-directional view maintains Causal Direction but rejects Temporal Direction. Consequently, each particular token of the causal relation still has a direction, but our world contains numerous causal relations that point in the backward direction in addition to the known causal relations in the forward direction.

Bi-directional causation holds that each token of the causal relation has a direction and is therefore different from a view where causation is a symmetric relation, which would also deny Causal Direction. For example, the relation *occurring two seconds apart* is symmetric. This relation lacks direction because for an event *c* to occur two seconds apart from *d* is the same fact as for *d* to occur two seconds apart from *c*.

In contrast, causation on my view behaves logically like the *loving* relation. Billy loving Suzy (luckily) allows for Suzy to also love Billy. Still, loving is not a symmetric relation because each token of loving is directed. Billy's love is directed at Suzy, and Suzy's love is directed at Billy. In a situation where Billy loves Suzy and Suzy loves Billy back, the lovers instantiate two distinct and oppositely-directed tokens of the relation. The one token is directed from Billy to Suzy; the other token is directed from Suzy to Billy. Moreover, the two tokens are distinct because each token consists in a different fact: Billy's loving Suzy consists in a mental state in Billy's mind and Suzy's loving Billy consists in a mental state in Suzy's mind.

Analogously, my bi-directional view of causation allows for situations where an event *c* causes another event *d*, and *d* also causes *c*. In these situations *c* and *d* instantiate two distinct and oppositely-directed tokens of the causal relation: one token points from *c* to *d* and one token from *d* to *c*. So each token of causation is directed and causation is bi-directional rather than time-symmetric.

Bi-directional causation thus leaves room for the time-asymmetry of causation because it allows that causal relations in the forward direction are different in character from causal relations in the backward direction. As an analogy, imagine a railway network that has trains running both from east to west and also from west to east. But if west-bound trains are

in general faster and more reliable than east-bound trains, then there is an important spatial asymmetry in the railway network despite the fact that trains run in both spatial directions. This asymmetry has important practical implications, making, for example, traveling westwards much easier than traveling eastwards.

Analogously, I argue that the perceived asymmetries associated with the arrow of causation are due to gradual differences between forward causation and backward causation. The two most important asymmetries associated with causation are control and explanation:

Time-asymmetry of control. We have some limited control over the future but absolutely no control over the past.

Time-asymmetry of explanation. Earlier events often explain later events but usually not *vice versa*.

I shall argue that causation can come apart from control and explanation, and that these asymmetries arise because, though causation goes in both temporal directions, causation in the backward direction lacks the very features that make causation in the forward direction suited for control and explanation. This divergence is important because an important reason for why backward causation strikes us as absurd is that we associate causation with control and explanation. But the backward causation that my theory entails neither gives us control nor does it support explanations.

4 Bi-directional causation and causal pluralism

What does it mean that causation is bi-directional? Many philosophers have observed that it does not seem that we have a uniform concept of *the* causal relation. Earman puts this point most poignantly, when he says that "causation is not a single, unary notion, but a multi-faceted concept which begs for distinctions to be drawn among various kinds of causal interaction." (Earman 1976b, 390)

Fortunately, my view that causation is in an important sense bi-directional is compatible with pluralism about our causal concept. Following Hitchcock (2003), we can think of theories of causation in two stages. The first stage of analysis "involves the identification of some privileged class of entity, and the discrimination of the members of this class from various impostors." (Hitchcock 2003, 5) The goal at this stage is to identify a non-accidental, directed dependence that can ground effective strategies and to distinguish it from mere correlation. We are familiar with this dependence from paradigmatic cases, such as when two billiard balls collide. Call this dependence "causal dependence."

As Hitchcock (2003) points out, the most common theories of causation can be seen as converging toward isolating causal dependence by trying to distinguish it from other relations (of the kind Hitchcock calls "impostors"). For example, counterfactual theories distinguish genuine counterfactuals from backtrackers; process theories distinguish causal processes from pseudo-processes; regularity theories distinguish lawlike from accidental regularities; and probabilistic theories distinguish real from spurious probabilities. All of these distinctions can plausibly be seen as isolating causal dependence.

My bi-directionality claim concerns this dependence relation that contrasts with mere correlation and that can ground effective strategies. I argue that the right way to delineate this

dependence is such that it goes in both temporal directions. We still have to distinguish causal dependence from impostors, but I argue that lots of dependencies in the backward direction are legitimate rather than impostors. This is a substantial metaphysical claim that most philosophers deny.

This view about causal dependence still leaves room for causal pluralism, which comes in at the second stage of analysis. The second stage concerns how these basic building blocks of causal dependence can be put together to make-up interesting relations. As Hitchcock points out, there is room for pluralism here because causal dependence can be put together in different ways. For instance, Lewis identifies causal dependence with counterfactual dependence but then goes on to identify causation with the ancestral of this relation, thus making causation a transitive relation (cf. Lewis 1986, 167). Other philosophers have denied that causation is transitive, and Hall (2004) argues that there are two distinct kinds of causation: one that is transitive and one that is not. Similar debates concern whether causation is intrinsic and how many relata it has.[10]

My view allows me to stay neutral on whether there is a single causal relation or whether there are several relations differing in these properties, as pluralists claim. I am only concerned with the basic building blocks of causation that are needed for a theory of effective strategies. None of the prominent arguments for causal pluralism suggest that there is pluralism at this level. Insofar as my thesis about bi-directional causation concerns the

[10] See Schaffer (2008) for an overview of these debates.

basic building blocks of causation, it is compatible with pluralism about our concept of causation. My thesis is only that there are more building blocks than we thought there were.[11]

5 Redrawing the arrow of causation

Widespread backward causation is contrary to our ordinary notion of causation. So it is not clear how a relation could be causal if it does not run at least predominantly in the forward direction. But drastic revisions of ordinary phenomena are familiar from other areas of science and metaphysics. I will look at two such cases to motivate the feasibility of a revisionary stance toward the direction of causation.

The first case concerns the nature of light.[12] Ordinarily, we think of light as the agent that makes things visible to us. However, there is a scientifically informed conception of light, where light covers all forms of electromagnetic radiation, even those that are not 'visible.' For instance, the dictionary lists the following as one definition of light: "electromagnetic radiation of any wavelength that travels in a vacuum with a speed of about 186,281 miles (300,000 kilometers) per second."[13] This conception revises our ordinary understanding because there is lots of electromagnetic radiation that does not make things visible to us. So physics shows that the familiar instances of light are a subset of a much broader phenomenon. The same agent that allows us to see things also comes in guises where it lacks the capacity to make things visible to us.

---

[11] My project is thus at least to some extent independent of semantic questions because "causal dependence" is a technical term picking out the metaphysically basic building blocks of causal relations that underlie effective strategies and that contrast with correlations.

[12] See Ney (2009, 760) for the analogy between causation and light.

[13] http://www.merriam-webster.com/dictionary/light.

The second case concerns the direction of time. On our ordinary conception, time has an intrinsic direction that grounds its transitory character. We think that present events continuously become past as they give way to future events. But many philosophers have argued that time has no intrinsic direction. These philosophers argue that time is one dimension of a four-dimensional manifold which has the same character in either temporal direction. Although this view of time drastically revises our ordinary conception of the nature of time, many philosophers and physicists take it extremely seriously.

The cases of light and time provide a blueprint for how to think of a revision of our understanding of causation in light of fundamental physics. There are important analogies between my views on causation and the case of light. Causal relations in the forward direction are familiar to us because they figure prominently in control and explanation. I argue that, just as there is light that is not visible, there are causal relations in the backward direction that are irrelevant to control and explanation. Our evidence for these causal relations is indirect and comes from the structure of the fundamental laws. As in the case of light, the revision does not concern the nature of the known instances of causation but merely shows that these instances belong to a broader class than we previously thought that also includes instances in the backward direction.[14]

The analogy to time is even closer. My argument for why causation is bi-directional closely parallels an argument in the literature for why time has no direction. Many of our best candidates for the fundamental physical laws are time-symmetric, which means that these

[14] Another such revision concerns background conditions. Ordinarily, we distinguish causes from background conditions. For example, we think that the striking of the match causes its lighting but the existence of oxygen is merely a condition. However, philosophers standardly concluded that there is no metaphysical distinction between causes and conditions. Both are equally causes

laws operate exactly the same way in the forward direction as they operate in the backward direction.

A popular argument says that it is reasonable to draw inferences about the nature of time from the structure of the fundamental laws.[15] After all, the fundamental physical laws fully describe all possible behaviors of systems in our universe. If the laws are time-symmetric such that the evolution of systems in the forward direction falls under the exactly same constraints as their evolution in the backward direction, then it is reasonable to assume that time, as the arena in which systems evolve, has the same character in both directions.[16] This argument is not indefeasible because there might be overriding reasons for thinking that time has an intrinsic direction, but the inference is reasonable (cf. North 2008).

My argument for bi-directional causation closely parallels this argument. Causation concerns how events evolve over time, and the fundamental physical laws completely describe how events at one time evolve into events at other times. If the laws are such that earlier events constrain later events in the exact same way as the other way round, then it is reasonable to think that there is no difference between how systems in our universe evolve forward and how they evolve backwards. It is thus reasonable to infer that causation runs both in the forward and in the backward direction. (See chapter 1 for further defense of this inference.)

My revisionary view of causation therefore has close precedents in other debates, and I argue that revision is as plausible in the case of the causal direction as in these other cases. I

[15] This inference is widely made in physics. For example, Greene (2005) takes the structure of the laws to indicate the nature of spacetime without even acknowledging that this step takes any kind of justification. See North (2008) for an explicit discussion of this argument.

[16] It does not matter currently whether the actual fundamental laws of physics are in fact completely time-symmetric because I am only concerned with the reasonableness of drawing inferences about the nature of time from time-symmetric laws.

think the main reason why philosophers have not seriously considered revisionary accounts of the temporal direction of causation is that directionality seems to be a more central feature of causation than the revised features in the two other cases.

As seen earlier, the main reason for endorsing objective causal facts is in service of a theory of effective strategies. Effective strategies, however, appear to be firmly temporally directed in that our actions are never effective strategies for earlier ends. So it seems that bi-directional causation could not ground effective strategies and thus would miss the point of a theory of causation. In response, I argue that bi-directional causation leads to a plausible theory of effective strategies. (I will only sketch the argument here, which is further developed in chapters 1 and 2.)

There is an important practical distinction between effective strategies that are *accessible* to agents like us and ones that are *inaccessible*. For example, taking antibiotics is an effective strategy for recovery from infection, and the strategy is also accessible because we can in fact bring about recovery via this course of action. In contrast, though decreasing the radius of a massive star is an effective strategy for creating a black hole, this strategy is inaccessible to us. The strategy does not in fact allow us to create a black hole because we cannot decrease a star's radius to the required extent.

One might object to calling these latter courses of action "effective strategies" because they are not action-guiding. But that is mere terminology.[17] The important point is that strategies, such as collapsing a star to create a black hole, are still grounded in causal relations. Decreasing the radius of a star would cause a black hole. It is merely that these

[17] Many normative concepts are ambiguous in that they allow for a *guiding* and an *evaluative* reading. For example, an agent's evidence can be understood evaluatively as any information the agent has that bears on the truth of a proposition, or it can be understood in a guiding sense as only the information that is accessible to the agent. The latter but not the former reading takes into account the agent's cognitive limitations.

causal relations do not have the same practical relevance for us because they do not allow us to control our environment. Moreover, we cannot directly experience or test these causal relations but have to infer them from observed data and the laws of nature.

I shall argue that the causal relations in the backward direction that my theory of bi-directional causation posits do ground effective strategies—but these effective strategies are inaccessible to agents like us. Though our actions cause past events, we cannot control past events. There are however important differences between the inaccessibility of backward-looking effective strategies and the collapsing star case. First, backward-looking effective strategies are inaccessible for a different reason. In chapter 2, I will argue that the inaccessibility is not due to a lack of muscle power or technological ingenuity (as in the star case) but due to lack of a certain kind of knowledge. Second, the inaccessibility of backward-looking effective strategies is more extreme than in the star case. In the collapsing star case, we can easily imagine agents relevantly like us to whom these strategies are accessible, such as technologically advanced humans, gods, or superheroes. Agents who could exploit effective strategies in the backward direction, in contrast, would have to be different from us in a more fundamental way. They would need a different cognitive make-up.

My theory of causation is thus compatible with a plausible theory of effective strategies. To begin with, it entails that there are objective distinctions between effective and ineffective strategies. It is objectively true that quitting smoking is an effective strategy toward avoiding lung cancer because smoking causes lung cancer. It is also objectively true that having your teeth whitened is not an effective strategy to avoid lung cancer because having yellow teeth does not cause lung cancer. Moreover, my theory can account for the feeling that it would be comical or futile to act for the sake of past ends by showing that

backward-looking effective strategies are in principle inaccessible to agents like us. So my theory still explains the facts about effective strategies that we care about, viz., why certain courses of actions are irrational and others are not. In addition, I will show that causation in the backward direction is irrelevant to the kinds of explanations the special sciences seek.

However, my theory is revisionary with regard to what metaphysical facts underlie these normative facts about our practices. I argue that backward-looking effective strategies are not impossible but merely inaccessible. Because of the time-symmetry of the fundamental physical laws, the distinction between past and future is metaphysically less deep than we ordinarily think. Just as we learn from the fundamental laws that, for example, decreasing the radius of a star would create a black hole, so we also learn that there are effective strategies in the backward direction.

This revision not only allows us to give a better account of the dependence structure of our universe but also enables us to give a naturalistic explanation of why we cannot control the past. Our inability to control the past is not due to some deep metaphysical difference between the past and the future. Instead, we lack this ability because of the character of the physical laws and our make-up as agents.

REFERENCES

Albert, D. (2000) *Time and Chance*. Cambridge: Harvard University Press.

Carroll, S. (2010) *From Eternity to Here: the quest for the ultimate theory of time*. New York: Dutton.

Cartwright, N. (1979) "Causal Laws and Effective Strategies," *Nous* **13**: 419-437.

Dowe, P. (2000) *Physical Causation*. Cambridge: Cambridge University Press.

Earman, J. (1976a) "Causation: a matter of life and death," *Journal of Philosophy* **73**: 5-25.

Earman, J. (1976b) "*The Cement of the Universe* by J. L. Mackie," *The Philosophical Review* **85**: 390-394.

Field, H. (2003) "Causation in a Physical World," in: Loux, M. and D. Zimmerman (eds.) *Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 435-460.

Frisch, M. (2005) *Inconsistency, Asymmetry, and Non-Locality*. Oxford: Oxford University Press.

Greene, B. (2005) *The Fabric of the Cosmos*. New York: Alfred A. Knopf.

Hall, Ned (2004) "Two Concepts of Causation," in: Collins, J., N. Hall, and L.A. Paul (eds.) *Causation and Counterfactuals*. Cambridge: MIT Press.

Hitchcock, C. (2001) "The Intransitivity of Causation Revealed in Equations and Graphs," *Journal of Philosophy* **98**: 273-299.

Hitchcock, C. (2003) "Of Humean Bondage," *British Journal for the Philosophy of Science* **54**: 1–25.

Lewis, D. (1986) "Causation," in: *Philosophical Papers: Volume II*. Oxford: Oxford University Press, 159-213.

Lockwood, M. (2005) *The Labyrinth of Time*. Oxford: Oxford University Press.

Maudlin, T. (2007) *The Metaphysics Within Physics*. Oxford: Oxford University Press.

Ney (2009) "Physical Causation and Difference-Making," *British Journal for the Philosophy of Science* **60**: 737-764.

North, J. (2008) "Two Views on Time Reversal," *Philosophy of Science* **75**: 201-223.

Norton, J. (2007) “Causation as a Folk Science,” in: Price and Corry (2007), 11-44.

Papineau, D. (1985) “Causal Asymmetry,” *British Journal for the Philosophy of Science* **36**: 273-289.

Price, H. (1996) *Time’s Arrow and Archimedes’ Point*. Oxford: Oxford University Press.

Price, H. and R. Corry (2007) *Causation, Physics and the Constitution of Reality*. Oxford: Oxford University Press.

Reichenbach, H. (1956) *The Direction of Time*. Berkley: University of California Press.

Russell, B. (1913) “On the Notion of Cause,” *Proceedings of the Aristotelian Society* **13**: 1-26.

Schaffer, J. (2008), “The Metaphysics of Causation,” *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), E. N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2008/entries/causation-metaphysics/>.

van Fraassen, B. (1993) “Armstrong, Cartwright, and Earman on Laws and Symmetry,” *Philosophy and Phenomenological Research* **53**: 431-444.

Weslake, B. (2006) “Common Causes and The Direction of Causation,” *Minds and Machines* **16**: 239-257.

Woodward, J. (2003) *Making Things Happen*. Oxford: Oxford University Press.

BLUNTING THE ARROW OF CAUSATION

Chapter One

1 Introduction

According to our ordinary view, causation has a strict temporal direction such that the world causally evolves forwards but not backwards. We think that the past brings about, produces, or shapes the future but not *vice versa*. But this forward-directed view of causation is in tension with how fundamental physics describes the world. In particular, most candidates for the fundamental physical laws determine the evolution of the universe in both temporal directions.[18] These laws determine the evolution of the world in the forward direction, but they equally determine its evolution in the backward direction.

Though we naturally interpret the laws as describing the causal evolution of systems in the forward direction, this interpretation is superimposed upon the laws, not derived from them. The fundamental physical laws describe how the state of the world at any one time depends on its state at other times. But nothing about these laws makes it any more apt to view the world as evolving forwards rather than as evolving backwards. Instead, earlier states lawfully depend on later states in the same way as earlier states lawfully depend on later

[18] Cf. Albert (2000, chapter 2), Carroll (2010, chapter 2), Field (2003, 436pp); Greene (2005, chapter 6); and Lockwood (2005, chapter 9).

states, and in this sense the fundamental laws determine the evolution of the world in both temporal directions. Let us call such laws *time-symmetric* laws.[19]

This time-symmetry is most familiar from the Newtonian laws, but it is equally part of contemporary laws such as Schrödinger's equation in quantum mechanics and the field equations in General Relativity. These laws are all deterministic in both temporal directions. That means, given the state of the world at any one time these laws fix a unique future *and* a unique past. Moreover, the problem does not significantly change if the laws are probabilistic as long as they have the same probabilistic character in both temporal directions. In this case, a complete specification of the universe at any one time, together with the laws, entails a probability distribution over all earlier and later times. These laws would still determine the evolution of the universe in either temporal direction by specifying probabilities.[20]

Much recent debate concerns how to reconcile causation with this bi-directional lawful evolution. One way of seeing the tension is by considering counterfactuals:

(C1) If the earlier momentum of the ball had been different, then its later momentum would have been different.

(C2) If the later momentum of the ball had been different, then its earlier momentum would have been different.

[19] I use "time-symmetric" for lack of a better word. Time-symmetry is also often used to describe *time-reversal invariance*, which means, roughly, that for every sequence of events which is in accordance with the laws, the time-reverse of that sequence is also in accordance with the laws. The relevant sense here, however, is that the laws connect earlier to later states in the same kind of way they connect later to earlier states (cf. Field 2003, 436). Time-reversal invariance is neither necessary nor sufficient for the laws having this feature.

[20] The fundamental laws could turn out to be only deterministic in the forward direction or only specify probabilities in the forward direction. In this case, the laws *would* single out a direction of temporal evolution. The only candidates for the fundamental laws which have this feature are some "collapse-theories" in quantum mechanics, which only specify probabilities in the forward direction. Collapse theories, however, are extremely contentious. Moreover, they have empirically equivalent competitors that are deterministic in both temporal directions (e.g., Bohmian Mechanics).

Causation and counterfactuals are closely related. Intuitively, *C1* is true but *C2* is false because the ball's earlier momentum causes its later momentum, whereas its later momentum does not cause its earlier momentum. But counterfactuals are also law-governed, and time-symmetric laws underwrite both *C1* and *C2*. Given that the ball is sufficiently isolated from its environment, a counterfactual state in which the antecedent of *C1* is true and the momentum of the ball is different lawfully entails a later state where its momentum is different. However, a counterfactual state in which the antecedent of *C2* is true equally lawfully entails an earlier state where its momentum is different.

Philosophers tend to respond to this puzzle in one of three ways. First, Compatibilists argue that our common-sense view of causation as forward-directed is compatible with bi-directional laws and bring in facts other than the laws to ground the direction of causation.[21] Eliminativists, in contrast, argue that our commonsense view of causation is incompatible with the fundamental physical laws and that causal relations are therefore not part of the objective physical world.[22] Both views, however, take our ordinary conception of the causal direction at face value and merely disagree on whether it latches on to anything in the physical world.

In this paper, I defend a new response to the problem: bi-directional causation. I argue that causation is compatible with fundamental physics but that we have to revise its temporal direction in light of the time-symmetry of the fundamental laws. Bi-directional causation is

[21] I will later in the paper distinguish reductive Compatibilist theories, which try to reduce the direction of causation to non-causal physical features, from primitivist theories, which that take the direction of causation to be primitive. Reductive theories include Dowe (2000), Field (2003), Lewis (1986a), Papineau (1985), and Reichenbach. Primitivists comprise Frisch (2005), Maudlin (2007), and Tooley (1987).

[22] See Norton (2007), Russell (1913), Price (1996, 2007), and van Fraassen (1993). Moreover, Neo-Russellians deny that causal concepts apply in fundamental physics, though they are legitimate in other contexts. See Hausman (1998), Hitchcock (2007), and Woodward (2007).

the view that causation runs both forwards and backwards. The causal arrow on this view is not strict, as causation also runs in the backward direction, but it is due to differences in character between causation in the forward and in the backward direction.

This view might sound absurd, but it follows from taking fundamental physics seriously. Rather than impose our ordinary experience and intuitions onto a theory of causation, we should develop our theory of causation in accordance with the structure and features given to us by fundamental physics, and only add features (like a privileged temporal direction) if there is some outstanding need to do so. I shall argue below that causation is law-governed and that time-symmetric fundamental laws ground causation in both temporal directions. Moreover, I will show that bi-directional causation is consistent with our experience.

My theory thus resolves the tension between causation and fundamental physics by revising the temporal arrow of causation. Comparably drastic revisions to our everyday understanding are familiar from other phenomena. For example, many philosophers argue that the perceived passage of time is merely a feature of our experience and that fundamental physics teaches us that time has no intrinsic direction. I will defend an analogous revision in the case of the temporal direction of causation.

I think that philosophers have not taken bi-directional causation seriously because it seems out of touch with our experience. If causation goes backwards, then, for example, it seems we should be able to control the past just as we can control the future. But I shall argue that causation can come apart from control as well as explanation, and that backward causation does not support these practices in the same way forward causation does. My theory of bi-directional causation therefore is compatible with our ordinary experience

because it allows for differences between the character of causation in the two temporal directions and hence for an important respect in which causation *is* time-asymmetric.

In the rest of the paper, I first argue that it is very natural to assume that causation is bi-directional if the laws are time-symmetric (Section 2). Second, I defend bi-directional causation against the main objections to the view (Section 3). Third, I show how bi-directional causation is compatible with our experience, in particular the time-asymmetries of control and explanation (Sections 4 and 5).

2 Time-symmetric laws and bi-directional causation

Causation and the laws of nature are closely related in that both concern how events evolve over time. Moreover, because the fundamental physical laws tell the complete and exceptionless story of how our universe evolves over time, it is natural to think that the structure of the laws has implications for causation. In this section, I argue that it is extremely plausible that causation is bi-directional if the laws are time-symmetric.

The fundamental physical laws determine how systems evolve over time by constraining, given the events at any one time, which events have to happen at other times. For instance, the position and momentum of a billiard ball at some time, plus a specification of all forces acting on it within some interval of time, together with the laws constrains its position and momentum over the interval.

Time-symmetric laws determine the evolution of the universe in both temporal directions: They determine its evolution in the forward direction, but they also determine its evolution in the backward direction. That is, given the state of a system at any one time, these

laws equally constrain both its earlier and later states. Time-symmetric laws therefore do not single out any direction in which the universe evolves.[23]

Think of the states of the world at different times as the frames of a movie. Given a full specification of any one frame, the laws fully constrain all later and earlier frames. And given a partial specification of any one frame, they partially constrain all other frames. The laws thus determine the evolution of the movie in both temporal directions by specifying how the content of any one frame constrains the content of all later and earlier frames.

Given time-symmetric laws, describing a sequence of events as a window breaking into pieces is no more apt than describing it as glass pieces forming a window. We can say, for example, that a stone hitting a window evolves forwards into glass pieces and a stone lying on the floor because the earlier events lawfully entail the later events. But we can equally say that the glass pieces and the stone lying on the floor evolve backwards into a stone flying away from an intact window because these later events lawfully entail the earlier events.

I will argue that, given this bi-directional lawful entailment, causation also goes in both temporal directions. This inference does not require that causation is identical to lawful entailment. In fact, there are good reasons for thinking that it is not. For instance, the complete state of the world *S1* at some time *t1* lawfully entails that there is a window shattering at some later time *t2*, but not all parts of *S1* are causes of the window shattering. Nonetheless, I will argue that there is a sufficiently close tie between causation and lawful

[23] For instance, if you use the formalism of Newtonian Mechanics to calculate the evolution of the universe toward a time of interest, you do not have to know whether this time is in the past or the future. You merely need to know all the fundamental quantities and how they change relative to the time of interest. Given this information, you would make the exact same calculations in either temporal direction.

entailment to make it plausible that if lawful entailment goes in both temporal directions, then causation also goes in both directions.

It is plausible that the structure of the fundamental laws determines the nature of causation, as the following cases illustrate. First, we justify causal claims by referencing the fundamental laws. If asked why its collision with another ball causes the billiard ball to move, it is natural to say that given the momentum of the incoming ball and the fundamental laws of physics the ball had to move. Moreover, we can use the fundamental laws to infer causal relations in circumstances where we cannot do experiments. For example, we can derive from the fundamental laws that the motions of the moon cause the tides because we can compute how changing the moon's orbit would change the forces on the tides and how the tides would behave given these forces.

Second, we take the structure of the laws to determine general features of the causal relation. For instance, we think that if the laws are chancy such that events lawfully constrain other events by fixing objective chances, then causation is also chancy, i.e., causes fix the objective chances of their effects (cf. Lewis 1986a). Moreover, it is plausible that if the laws were non-local, then causation would be non-local. Laws are temporally non-local, just in case: possibly, there is some event *d* whose occurrence is lawfully determined by the events at some time *t*, but there are no events that lawfully determine the occurrence of *d* at some time *t** that is in-between *t* and the time of *d*. In other words, temporally non-local laws act across a temporal gap. It is reasonable that in such circumstances causation would also be non-local in that *d* would lack intermediate causes at *t**.

These cases establish that it is natural to assume that the structure of the fundamental physical laws determines both particular causal relations and global features of the causal

relation. Getting clearer on the nature of this determination will show that it is equally plausible that the fundamental laws determine the temporal direction of causation. After all, the temporal direction of causation concerns both particular causal relations (viz., whether there are any in the backward direction) and a global feature of causation (viz., its temporal orientation). At least pre-theoretically, we expect that the direction of causation is written into the fundamental laws of physics.

The defining feature of causes is their efficiency: causes make their effects happen. That is, given the causes, the occurrence of the effect is not an accident; the effect has to occur.[24] For instance, given that the stone hits the window (and the circumstances), the window has to shatter. But what grounds this entailment? Why do effects *have* to occur given their causes?

It is extremely plausible that causes entail their effects because of facts about lawful entailment. Given that the cue ball bumps into it, the eight ball has to move because given the collision, and suitable background conditions, the fundamental physical laws entail the moving of the eight ball. It is unclear how else to understand causal efficiency. A complete specification of a physical system plausibly incorporates three elements: a description of its actual state at the time, static laws governing what possible states it could be in, and laws of temporal evolution. Of these elements, the laws of temporal evolution are the only entities that are relevant to how events at one spacetime region evolve into events at another spacetime region. So it is extremely plausible that causal efficiency consists in facts about

[24] I am assuming that the laws are deterministic, but the claim could be rephrased to take into account indeterministic laws. In that case, causes make their effects happen such given the causes, the occurrence of the effect is non-random, but the effect has to have a certain objective chance.

lawful entailment. Furthermore, causal efficiency is both necessary and sufficient for causation. An event *c* causes *d*, just in case *c* makes *d* happen.

The claim that causation holds in virtue of facts about lawful entailment leaves open important questions about how exactly relations of lawful entailment ground causation. As said, it is not plausible that lawful entailment is identical to causation. But for present purposes, we can set these details aside. How exactly lawful entailment grounds causation will not matter for the temporal direction of causation because lawful entailment works the same in both temporal directions. So if it grounds causation in the forward direction, it is equally plausible that it grounds causation in the backward direction.[25]

Take an event such as the collision between two billiard balls at some time *t1*. There are events at some time *t0*, a few milliseconds earlier, that lawfully entail the collision, such as the positions and momenta of the two balls. These events are earlier causes of the collision. But if the laws determine the evolution of systems in both temporal directions, then there are also events at some time *t2*, a few seconds later, that lawfully entail the collision of the balls at *t1*. For example, the later positions and momenta of the ball at *t2* also lawfully entail their earlier collision. For that reason, the same kinds of facts about lawful entailment obtain in both temporal directions, and it is therefore plausible that they ground causal relations in the backward direction just as they do in the forward direction.

[25] Though causes bring about their effects in virtue of facts about lawful entailment, we do not regard all the events that lawfully entail an effect as causes. Intuitively, only events are causes that play a special role in the entailment of an effect in the sense that they are responsible for the fact that the laws entail this particular effect. Theories of causation try to single out the causes, for example, in terms of minimal sufficiency, counterfactuals, probabilities, or transferred quantities, which are in turn determined by facts about lawful entailment. (See Hall 2005 for discussion.) But however this issue gets resolved, because the laws work the same way in both temporal directions, whatever facts about lawful entailment ground causation in the forward direction equally obtain in the backward direction.

Though unfamiliar, it is in fact rather natural to describe these later events at *t2* as causes. After all, they make a difference to the earlier collision. Given the positions and momenta of the balls at *t2*, the collision at *t1* has to occur. But had the position and momentum of one ball or both balls been different, the laws would entail that a different collision or no collision would have occurred. Later events thus make a difference to earlier events in that different later events would lawfully entail different earlier events. It is extremely plausible to think of causation as a difference-making relation. Hence, it also comes natural to say that the collision depends on these later events; that these later events are responsible for the earlier collision; and that the collision happened *because* of these later events.

Such relations between later events and earlier events have all the features we typically associate with causation, apart from temporal precedence. Imagine watching a movie of the billiard game that is run in reverse, and suppose you see a sequence where two balls are colliding. Seeing the eight ball bumping into the cue ball, you would have no trouble in describing this interaction as causal and to identify the momentum of the eight ball as a cause of the momentum of the cue ball. The interaction appears to have all the features that we typically associate with causation: it is spatiotemporally contiguous, momentum is transferred, the cue ball's movement counterfactually depends on the eight ball's movement, and the eight ball's movement makes the cue ball's movement more probable. You would not even notice that the movie is played in reverse if you see just this sequence.

Moreover, it seems that, upon discovering that the movie is run in reverse, nothing would undermine your causal judgment. You might think that the interaction is not causal because the alleged cause does not temporally precede its effect. It is, however, hard to see

how temporal precedence by itself could make such a difference, especially because the fundamental physical laws treat evolution in the two temporal directions exactly the same. Some philosophers have argued that such interactions are not causal because of relations the eight ball's and the cue ball's movement bear to other events, such as the heating up of the table and the movements of air molecules. However, it is implausible that causation is extrinsic in this way.[26]

For these reasons, it is very natural to assume that if the laws are time-symmetric, then later events cause earlier events just as earlier events cause later events. I will further defend this bi-directional view of causation in the next section. But first I will look at how bi-directional causation revises our ordinary understanding of the temporal direction of causation.

We can separate out two ways in which we ordinarily regard causation as directed:

(i) Causal Direction. Causal relations are directed from cause to effect (i.e., *c* causing *d* is different from *d* causing *c*).
(ii) Temporal Direction. Causes often precede their effects, but effects do not (or at least not typically) precede their causes

Causal Direction says that each particular token of causation has a direction. It does, however, not entail that causation is time-asymmetric. For example, if all cause-effect pairs were simultaneous, tokens of causation could still be directed but causation would not be

[26] I have in mind here views like Lewis (1986a) and Papineau (1985) that tie causation to extrinsic facts such as overdetermination (Lewis) or forking (Papineau). I will criticize these proposals in more detail in the next section.

time-asymmetric. The common-sense view of the causal direction thus, first, requires individual instances of causation to be directed; and, second, it requires causes and effects to be distributed in time such that causes (typically) precede their effects.

My bi-directional view maintains Causal Direction: Each particular causal relation has a direction. But it denies Temporal Direction: Our world contains, in addition to the known causal relations in the forward direction, also numerous tokens of causation that point in the backward direction. So causation is widespread both in the forward and in the backward direction of time.

Because it maintains Causal Direction, causation on the bi-directional view is not a symmetric relation. Tokens of symmetric relations, such as *occurring two seconds apart*, lack direction. For instance, for *a* to occur two seconds apart from *b* is the same fact as for *b* to occur two seconds apart from *a*. On my view, each token of the causal relation has a direction, and so causation is not a symmetric relation. If the stone hitting the window causes the shattering, and the shattering also causes the stone hitting the window, then the two events instantiate two distinct and oppositely directed tokens of the causal relation. The one causal relation is grounded in lawful evolution in the forward direction; the other causal relation is grounded in lawful evolution in the backward direction. These two facts are logically distinct. Similarly, when Billy loves Suzy and Suzy loves Billy, there are two tokens of the loving relation pointing in opposite directions.

This fact is important because it allows for causation to still be time-asymmetric despite running in both temporal directions. On the common-sense view, causation is *strictly* time-asymmetric: causation is present in the forward direction but absent (or at least

extremely rare) in the backward direction. But causation can be time-asymmetric without being strictly time-asymmetric, as the following analogy illustrates.

A railway network that operates trains only westwards is strictly spatially asymmetric. But a network can be asymmetric even if it operate trains both east- and westwards because there can be qualitative differences in the two directions. West-bound trains might, in general, be more reliable and faster that east-bound trains, making it much easier to travel west than east. Analogously, I will argue that causal relations in the backward direction are different in character from causal relations in the forward direction, which makes for a drastic practical difference in their availability for control and explanation.

I thus argue that causation is *qualitatively* time-asymmetric: there are numerous tokens of causation pointing forwards and numerous tokens pointing backwards, but there are qualitative differences between causation in the forward and in the backward direction. (I will point out some important differences in sections 4 and 5.) So rather than denying that causation has a temporal direction my view re-conceives what that direction is in a accordance with the time-symmetry of the fundamental physical laws.

3 Bi-directional causation defended

I have argued that it is extremely natural to develop our theory of causation in accordance with the structure of the fundamental physical laws and that the resulting view of causation is one where causation is bi-directional. In this section, I will further defend this bi-directional theory of causation.

There are three main objections to bi-directional causation. The first objection says that bi-directional causation is untenable as a view of causation. Causation is intimately

bound up with a number of important time-asymmetric practices. So the charge is that if there were massive backward causation, we would lack a reasonable account of the time-asymmetries of these practices. It seems we should then be able, for example, to control the past and causally explain earlier events by citing their later causes, which is absurd.

Objection from practical relevance. Bi-directional causation is incompatible with important practical time-asymmetries that are associated with causation, such as control and explanation. Therefore, bi-directional causation is untenable.

In reply, I will argue that my bi-directional theory of causation is compatible with these time-asymmetric practices. In sections 4 and 5, I show that my bi-directional view can account for the two most prominent time-asymmetries, viz., control and explanation. Because we typically associate causation with control and explanation, my theory can explain why we overlook causation in the backward direction and thus think that causation goes only forwards. Backward causation is under our radar because it is irrelevant to control and explanation.

The second objection says that my account ignores important physical asymmetries. Compatibilist theories of causation hold that causation is forward-directed even if the laws are time-symmetric because the direction of causation is grounded in asymmetries other than the laws. These theories typically hold that causation has a time-symmetric component that is determined by the laws, but they hold that it additionally has another component that is time-asymmetric.

Objection from Compatibilism. Bi-directional causation focuses exclusively on the fundamental laws of temporal evolution but ignores the relevance of other physical asymmetries to the direction of causation. The forward-direction of causation, however, is determined by time-asymmetries from these other sources, which are compatible with the time-symmetry of the fundamental laws. Therefore, bi-directional causation is false.

Compatibilist theories deny that causation is determined by the structure of the laws in the way I have defended in the last section. Call this claim they reject "Law-governedness."

Law-governedness. The structure of the fundamental physical laws determines the nature of causation.

I will respond to the objection by defending Law-governedness. We can see the plausibility of Law-governedness by considering an analogous principle for the nature of spacetime. In constructing a theory of spacetime, it is reasonable to assume that its nature is determined by the structure of the fundamental physical laws. For instance, if the fundamental laws are completely time-symmetric (such that they treat past and future entirely on par), then spacetime itself has no temporal direction. This inference is reasonable because the fundamental physical laws completely specify the possible behavior of systems in spacetime. So if the laws constrain the behavior of systems in exactly the same way in each temporal direction, then we can reasonably conclude that spacetime itself has the same character in each temporal direction.

This inference is not indefeasible as there might be overriding reasons for adding a temporal direction to spacetime. For instance, an intrinsic direction of time might be needed to account for our experience of time's passage. The argument shows, however, that it is reasonable to construct our theory of spacetime in accordance with the structure of the fundamental physical laws and only add further features *if* there is some outstanding need for them.[27]

Law-governedness is the analogous claim for causation. The fundamental physical laws fully specify how events at one time evolve into events at other times. So we should develop our theory of causation in accordance with the structure of the laws and only add further features if there is some outstanding need to do so. In particular, if the laws determine the evolution of events the same way in each temporal direction, then we can reasonably assume that causation works the same way in each direction. This inference from the character of lawful determination to the nature of causation is reasonable because (as shown earlier) there is a clear connection between the physical laws and causation. Causation and the laws both concern why specified events at one time evolve into particular events at a different time. Moreover, the fundamental physical laws tell us all the facts about this evolution; otherwise, they would not be complete. So it is hard to see what other facts should matter to the character of causation.

Compatibilists, in contrast, deny Law-governedness by holding that other features, besides the laws of nature, also determine the nature of causation. Most Compatibilists hold that these features come from the boundary conditions. In physics, the total state of the world

[27] North (2008) explicitly defends this argument. Moreover, the argument is widely endorsed by physicists. Greene (2005) is representative of many in that he assumes that the structure of spacetime matches the structure of the fundamental laws without even acknowledging that this inference is in need of justification. See also Ismael (2011).

is accounted for by the fundamental laws together with temporal (and on some theories spatial) boundary conditions. Given the boundary conditions, the laws entail the state of the world at all other times.

Even if the fundamental laws are time-symmetric, there can still be time-asymmetries due to the boundary conditions. The dominant time-asymmetry in our universe is that many types of processes happen in the forward direction but never in the backward direction. For instance, apples fall from trees, but they never spontaneously jump upwards and fasten themselves to tree boughs; windows shatter, but shards on the floor never form windows; humans grow older but never younger; cigarettes burn to ashes, but ashes never reconstitute cigarettes, etc. These so-called "irreversible processes" are associated with an increase in entropy. The Second Law of Thermodynamics says that entropy in our universe never decreases and typically increases toward the future.

Most physicists think that the thermodynamic asymmetry is grounded in the boundary conditions (cf. Albert 2000 and Carroll 2010). In addition, there are several other systematic time-asymmetries in our universe that are closely related to the thermodynamic asymmetry. Many Compatibilists have argued that the direction of causation is grounded in such time-asymmetries from the boundary conditions, such as independence (Hausman 1998), forking (Dowe 2000, Papineau 1993, and Reichenbach 1956), or overdetermination (Lewis 1986a).

I will argue, however, that these asymmetries are of the wrong kind to determine a forward-directedness of causation. For instance, the Second Law of Thermodynamics describes *how* the world evolves in each temporal direction. It entails that systems typically evolve into states of higher entropy in the forward direction, but typically evolve into states of lower entropy in the backward direction. The world evolving differently forwards than

backwards, however, is different from the world evolving *only* forwards. The Second Law merely says that certain sequences of events in the forward direction (entropy-increasing ones) are incredibly more likely than other sequences (entropy-decreasing ones). But there is nothing inherent in entropy-increase that suggests that it, rather than entropy-decrease, should be associated with the direction of causation. Therefore, the fact that there is an entropy-gradient gives us no reason to think that causation goes forwards but not backwards.

This point generalizes to all asymmetries from the boundary conditions. For some feature to ground the forward-directedness of causation, its presence in the forward-direction would need to explain why earlier events make later events happen, and its absence in the backward-direction would need to explain why later events do not make earlier events happen. However, we can seamlessly make sense of causation even in situations where asymmetric features that arise from the boundary conditions are absent. So the absence of these features in the backward direction cannot explain why causation does not run backwards since we can equally make sense of causation without these features.

One such case concerns microscopic interactions that involve only very few particles, such as a collision between two electrons. Time-asymmetries in the boundary conditions do not show up in such cases, yet we think of such interactions as causal (cf. Price 1996, 151). Another case concerns hypothetical worlds that manifest no asymmetries in the boundary conditions at all, such as a world that only contains two colliding electrons. Again, we think that such worlds can contain causal interactions despite the absence of asymmetric boundary conditions (cf. Tooley 1987, 227).

In response, Compatibilists might argue that causation is only a macroscopic phenomenon (cf. Field 2003), or restrict their theories to the actual world (cf. Papineau 1985

and Dowe 2000). But my objection is not that Compatibilism fails to reproduce our intuitions in microscopic or hypothetical cases. Rather, the objection is that Compatibilism cannot explain why causation goes only forwards even on the macroscopic level in the actual world. If causation makes sense in the absence of features such as forking or entropy-increase, then the actual absence of these features in the backward direction cannot explain why causation does not go backwards in the actual world even on the macroscopic level. So while there is a clear story of how the fundamental laws matter to the temporal direction of causation, there is no equally compelling account of how the boundary conditions could determine a temporal direction of causation.

I think Compatibilist theories are popular, despite this shortcoming, for two reasons. First, many philosophers of causation engage in conceptual analysis and try to isolate a relation that is co-extensive with our concept of causation. We think that causation goes only forwards, and because, for instance, entropy also increases in the forward direction, it is tempting to identify the direction of causation with the direction of entropy-increase. However, part of the intuitive causal asymmetry is that the past determines the future in a deeper or more important sense than *vice versa*. Not just any asymmetry can underwrite this difference; and, as I have argued, asymmetries from the boundary conditions cannot.[28]

Second, time-asymmetries in the boundary conditions are connected to practical asymmetries that are associated with causation, in particular control and explanation. So it seems plausible that if we explain, for example, why we cannot control the past, then we have also explained why causation does not go backwards. And we can in fact account for

[28] For example, Lewis (1986a) and Dowe (2000) both identify the direction of causation with some physical asymmetry that is allegedly co-extensive with our ordinary causal concept without motivating how this asymmetry is supposed to vindicate our other beliefs about causation. See Horwich (1986, p. 171ff) for criticism of this feature of Lewis's account. See also Woodward (2003, chapter 3).

the time-asymmetries of control and explanation by bringing in asymmetries from the boundary conditions (see sections 4 and 5).

Such accounts, however, fall short of vindicating our ordinary view of the temporal direction of causation. Our ordinary belief that causation goes forwards but not backwards is not exhausted by the idea that we can control the future but not the past. We also think that this causal asymmetry has a principled character, specifically, that there is some sense in which the world metaphysically unfolds in the future direction that makes it in principle impossible to control the past. But I shall argue that the boundary conditions do not vindicate such a principled difference; they only support the idea that *we* cannot control the past due to limitations of our agency. This difference in degree, however, fits best with my bi-directional theory of causation.

The two most developed theories of the time-asymmetry of control and explanation fit my bi-directional view of causation better than they fit our ordinary view of causation. Hausman (1998) argues that "[w]hat characterizes causation and causal explanation is a certain *modularity*, which permits us to factor out influences and give us reason to pick out some nomological relations, to dub them <<causal,>> and to use them asymmetrically in explanations."[29] (Hausman 1998, 232; italics in the original) Hausman does not hold that influence runs only in the forward direction; influence in the forward direction is merely better behaved than influence in the backward direction, which makes it suited for control and explanation. Similarly, according to Albert (2000)'s influential account it is objectively true that the past does depend on the future in the causal sense of "depend," but agents like us

[29] Hausman 1998, 232; italics in the original. That Hausman does not think of the direction of causation as a deep metaphysical fact is also clear from when he suggests that the "thought that causes *necessitate* their effects is a metaphysical pun on the fact that agents "make" their effects happen by means of their causes." (Hausman 1998, 96; italics in the original)

cannot use this dependence in the past direction for control.[30] Both views thus take important steps towards a bi-directional theory of causation, where we think of the causal asymmetry as concerning the quality of causal relations rather than their absence in the backward direction (as illustrated by the train analogy in section 2).

This view has an important analog in the philosophy of time. Philosophers who deny that time itself has a direction still think that there are important asymmetries *in* time, concerning how material processes are arranged. For example, people's births always precede their deaths in the same direction of time. These theories then use asymmetries in time to account for our temporal experience while maintaining that these asymmetries do not vindicate our ordinary belief that time itself has a direction. Similarly, I argue that bi-directional causation explains the practical asymmetries associated with causation, but it does not vindicate a deep metaphysical difference in kind between past and future.

A final objection says that even if I am right that the closest relation to causation we find in the physical world is bi-directional, this fact would not show that *causation* is bi-directional. Instead, it would show that there cannot be a physical account of causation.

Objection from elimination. Forward-directedness is an essential feature of causation. So any relation that is bi-directional cannot be the causal relation.

[30] See Albert 2000, chapter 6. Barry Loewer, Albert's collaborator, explicitly distinguishes “influence,” which goes in both temporal directions, from control, which goes only forwards (Cf. Loewer 2012, 127ff).

Eliminativists argue on these grounds that there are no objective causal facts. Primitivists, in contrast, argue that we need causal relations and we therefore have to posit them as primitive entities.[31]

In response, I argue that my bi-directional relation is still recognizably a causal relation. In particular, my view maintains the central objective distinction between causation and mere correlation. For instance, the earlier air pressure causes the later occurrence of the storm, whereas the occurrence of the storm and the barometer reading are merely correlated as effects of a common cause. My view retains the objectivity of this distinction, and merely adds that there are more instances of causation than we thought, where the additional instances point in the backward direction. Moreover, it allows for causation to play a central role in control and explanation, as I will show in the next section. So Eliminativism is an overreaction because time-symmetric laws allow for causal facts, and Primitivism is not needed because, as I shall argue, we can account for the time-asymmetries associated with causation without primitive causal facts.

4 The time-asymmetry of control

The main objection against bi-directional causation is that it seems incompatible with the time-asymmetry of control.

The objection from control. Control has a temporal arrow. That is, we have some limited control over the future but absolutely no control over the past. However, if causation is bi-

[31] Norton (2007), Russell (1913), Price (1996, 2007), and van Fraassen (1993) defend Eliminativism. Frisch (2005), Maudlin (2007), and Tooley (1987) defend Primitivism.

directional and thus widespread in the backward direction, we should be able to control the past. Therefore, bi-directional causation cannot account for the arrow of control.

I will argue that the arrow of control is compatible with my bi-directional view of causation.[32] Control, in the sense relevant to agency, is an agent's ability to bring outcomes in accordance with her desires. Control in this sense can come apart from causation. Here is an example where an agent lacks control over an outcome even though her decisions cause the outcome:

Hurricane. A platitude from Chaos theory says that my current decision to clap my hands can cause a hurricane in Chile six weeks from now. An otherwise identical state of the world, except without me clapping, would not lead to a hurricane. But I still cannot control hurricanes in Chile.

My clapping causally influences the atmospheric conditions around the globe that then cause a hurricane in Chile six weeks later. Yet, I lack control over whether a hurricane will hit Chile.

I lack control because I cannot know, at the moment of my decision, whether my decision to clap would make the hurricane objectively more likely than my decision not to clap. For all I know, the circumstances might be such that my decision does cause a hurricane. But the circumstances might also be such that a hurricane is poised to happen

[32] A number of other accounts of the arrow of control in the literature are equally compatible with causation being bi-directional. See Albert (2000), Loewer (2007), and Price (1996, 2007). Though these authors hold a different metaphysics of causation, the resources they use to account for the arrow of control equally fit with my view of causation.

otherwise and my clapping would prevent it. Or, there might be no hurricane either way. Which scenario I am in depends on very fine details of the circumstances, in particular, the air currents all around the globe. But I cannot know these details, and so I cannot know, at the moment of my decision-making, which scenario I am in; and so I cannot know whether my decision to clap would make the hurricane more or less likely. So the case shows that control requires, in addition to causal influence, a certain kind of knowledge. I lack control over the hurricane because I cannot *know how* to bring it about.

Cases like *Hurricane*, where we lack control despite causal influence, are widespread. Our decisions cause numerous outcomes (in particular, microscopic events and events in the distant future) where we cannot know in advance whether or not a given decision makes the outcome more or less likely. For instance, my current finger movements might cause some air molecule to hit a certain spot on the wall, but I have absolutely no control over that outcome because I cannot know in advance whether or not my decision to move my fingers will cause that outcome. Therefore, it does not make sense for me to move my fingers for the sake of the outcome.

Causation is thus compatible with a lack of control. Every theory of causation needs to explain why we lack control in some cases but not in others. I will argue that we lack control over the past, despite causing past outcomes, for the same reason that we lack control in cases such as *Hurricane*. We cannot know how to bring about particular past outcomes. So there is a principled, non-ad hoc story of why bi-directional causation does not allow us to control the past.

In *Hurricane*, we lack control partly because of facts about the causal relation, but also, and importantly, partly because of facts about us. The fact about causal relations that matters to control is that they can be more or less sensitive or robust in the following way:

A causal relation from *c* to *d* is **sensitive**, just in case: there are no close variations of *c* or of the background conditions (or only very few variations), such that *c* would still cause a macroscopically similar outcome to *d*.

A causal relation from *c* to *d* is **robust**, just in case: there are many close variations of *c* or the background conditions such that *c* would still cause a macroscopically similar outcome to *d*.

The sensitivity of a causal relation is thus a measure of how much the circumstances can be varied while still causing a certain type of outcome. The more sensitive a causal relation is, the more specific the circumstances need to be chosen for it to obtain; hence, the more knowledge an agent would need to use it for controlling the effect.

In paradigmatic cases of control, our decisions cause the outcome robustly. For instance, my decision to throw a stone at a window causes the shattering of the window irrespective of the exact details of my decision or the background conditions. If I throw at a slightly different angle; if I pick a slightly heavier stone; if the wind is slightly stronger; or, if the weather in Chile is radically different, then the window still shatters.

This robustness is important because when we make decisions we are ignorant of many details of both the background conditions and our own decisions. When I deliberate

about whether to throw a stone or not, I have only approximate knowledge about what the world would be like given my decision. For instance, I do not know how the decision would be microscopically realized in my brain, the exact state of my muscles, or the exact weight of the stone in my hand, and so on. But because the causal relation is robust, my decision would cause a shattering for a wide range of ways in which these details could turn out. So despite my ignorance, I can typically know whether a certain decision would cause the outcome.

In contrast, if causal relations are extremely sensitive, then an agent needs a lot more knowledge to exploit them for control. For instance, knowing whether some decision of mine causes a hurricane in Chile six weeks from now would require extremely detailed knowledge about the background conditions and my own decision. Such sensitive causal relations allow for control in principle. A creature like a Laplacian demon, who has complete knowledge of what the world would be like given each available decision and has also unlimited calculating power, could utilize causal relations for control no matter how sensitive they are. However, for agents like us there are limits to how much we can know, and so if causal relations are very sensitive, like in the hurricane case, we cannot use them for control even in principle. Control thus depends both on how sensitive a causal relation is and how much we can know about the circumstances of our decisions; and it breaks down when the sensitivity of causal relations exceeds our knowledge.

If I can show that our decisions cause past outcomes only extremely sensitively, then I have shown that my bi-directional view entails that we are unable to control the past. I will argue that the fact that our decisions cause past outcomes only very sensitively is a consequence of our make-up as agents. Our decisions are small, localized events in our

brains.[33] Moreover, in intentional action these brain events always occur as parts of distinctive cognitive and physiological processes. Roughly, we receive information through our senses, deliberate and weigh options based on this information, make a decision, and then execute and consequently remember the decision. Physiologically, we can think of such processes in terms of electromagnetic and pressure waves reaching our sense organs, which trigger processes in our brains that send electric signals to our spinal cords, which then activate action potentials that lead to muscle contractions.

A remarkable fact about this hard-wiring is that our decisions get robustly 'magnified' to the macroscopic level in the forward direction by causing the future positions of, for instance, our arms and legs. That is, the processes by which our decisions cause the future positions of our body parts are extremely robust against changes in the environment. For example, my decision to lift my arm causes my arm to go up, largely irrespective of what I have had for dinner, the air pressure in the room, and other facts about my environment and myself. This robustness allows us to know the future effects of our decisions.

Without this robust 'magnification' of our decisions to the macroscopic level, our decisions would not allow us to control the future. To see this, imagine a person where all neural connections between her brain and the muscles in her body are blocked. Her decisions (thought of as brain events) would still cause future outcomes. For instance, her brain state matters to the temperature of her head, which in turn influences the movements of the air molecules in the immediate vicinity of her head. But her decisions no longer cause any macroscopic outcomes *robustly*. And so despite her causal influence on the future, we would

[33] In the following, I assume that our decisions are brain states. However, even a dualist who denies this assumption will grant that our decisions are realized by brain states or at least interact with the physical world via brain states, which would still allow me to make my point.

describe her as “trapped” in her body, unable to control her environment. Due to the damage in her hard-wiring, her decisions cause future outcomes only very sensitively, and so she is unable to control the future

We are in the very same situation with respect to the past. Given how we are hard-wired, our decisions do not robustly cause any past outcomes. We can see this by running our decision-making processes backwards. Viewed in the backward direction, the decision events in our brains influence the signals that connect them to our sense organs; and what is happening in our sense organs influences the details of the electromagnetic and pressure waves that leave our sense organs. So the only immediate past effects of our decisions are small differences in these outgoing waves. These differences, however, are very small and subtle. Of course, they might (and probably will) cause big differences in the more distant past. But the nature of these big differences depends on very sensitive features of the environment, and so we cannot know whether any of our decisions would make any past effect more or less likely. Just as the paralyzed person cannot know which future outcomes her decisions cause, we can never know which past outcomes our decisions cause, and so we lack control over the past.

But why are we hard-wired such that our decisions cause some future outcomes robustly but past outcomes only sensitively? A pervasive feature of our universe, due to the boundary conditions, is that macroscopic processes are extremely sensitive in the backward direction. Small changes to the state of a system typically lead to the same macroscopic future but lead to a completely different macroscopic history of the system. For instance, consider a stone that hits a window and shatters it. We can imagine many small changes to the trajectory of the stone that would not alter its macroscopic future. For example, if the

stone were a bit heaver or a bit faster, the window would still shatter. As said, this robustness is the basis of why we can shatter windows by throwing stones.

However, the same process is extremely sensitive in the backward direction. Because it is hard to picture processes in the backward direction, I will instead consider the time-reverse of the process, that is, a process where glass pieces on the floor jump upward and form an intact window. The particles in this reverse process go through exactly the same motions in the forward direction that the particles in the original process go through in the backward direction. Hence, if this time-reverse process is sensitive in the forward direction, then the original process is sensitive in the backward direction.[34]

The time-reverse of a window shattering is the following process:

Reverse. Glass pieces lie on the floor; at a certain moment the glass pieces and the stone first begin to vibrate and then move along the floor in little jumps that become increasingly larger. At the same time air waves from around the room converge inwards toward the glass pieces. Finally, the pieces and the stone jump upwards at the same time in coordinated movements; the glass pieces collide with each other, as the stone passes through, and chemically bond to form a smooth glass surface that fills the wall-opening.

Reverse is extremely sensitive because it only leads to an intact window in the future due to extreme coordination between its sub-processes. If a few particles in the floor move differently, they will not hit the glass pieces in the right way to make them jump upwards;

[34] This sensitivity, as well as this way of illustrating it, is beautifully explained in Elga (2000) who is building on work from Albert (2000). Elga uses this sensitivity as a counterexample to Lewis (1986a) but does not further discuss its significance.

and if just a few glass pieces move differently, they will not form a window. Small changes in the earlier conditions thus lead to a different macroscopic future, which means that it is true of the original, non time-reverse process that small changes in the later conditions lead to a different macroscopic history of the system. Window shatterings thus manifest a time-asymmetry of sensitivity.

Moreover, these processes are typical in this respect. It follows from standard thermodynamics that macroscopic processes in our universe are fairly robust in the forward direction but extremely sensitive in the backward direction.[35] The arrow of control is thus a consequence of a more general temporal asymmetry, viz., the time-asymmetry of sensitivity, which is closely related to the thermodynamic asymmetry. This time-asymmetry delivers the promised qualitative difference between causation in the forward and causation in the backward direction. Some macroscopic causal processes in the forward direction are robust, but all causal processes in the backward direction are extremely sensitive.

The time-asymmetry in our make-up as agents is thus an instance of this general asymmetry. We are ourselves macroscopic systems that are subject to the Second Law of Thermodynamics and our decisions-making processes are irreversible. So like all other thermodynamically irreversible processes, we expect them to be extremely sensitive in the backward direction.

In sum, despite backward causation, we lack control over the past. We can imagine creatures more knowledgeable than we are who know exactly which past and future outcomes each of their decisions cause. Such creatures could utilize backward causation for

[35] The precise explanation is that only a tiny subregion of the region in phase space taken up by the state of a macroscopic system corresponds to a macroscopic history that we regard as normal. In contrast, the overall majority of such a region corresponds to macroscopic futures that we regard as normal. See Albert (2000).

controlling the past because they would know what decision to make in order to bring about a particular past outcome. But given our make-up and cognitive capacities we lack this ability.

5 The time-asymmetry of explanation

Another objection against bi-directional causation appeals to explanation.

The objection from explanation. Explanation has a temporal arrow. In science and everyday life, we cite earlier events to explain later events but not *vice versa*. Moreover, many (perhaps all) explanations cite causes. But if causation were bi-directional, then citing later causes to explain earlier events would be just as valid as the other way round. Therefore, bi-directional causation is incompatible with the arrow of explanation.

In response to the objection, I argue that backward causation has no objectionable consequences for explanation. My view can account for why scientific explanations cite only earlier causes and why we feel cognitive relief when hearing about the earlier causes of an event but not when hearing about its later causes. In particular, I will argue that explanations that cite causal relations in the backward direction are practically uninteresting and also lack important objective virtues of good explanations.

I will again establish this point indirectly by considering the time-reverse of an ordinary causal process. If I show that this process in the forward direction does not support compelling explanations, then I have indirectly shown that the original process in the backward direction does not support compelling explanations, and hence how backward causation is compatible with the absence of explanation.

As an example, go back to *Reverse*. We would be shocked to see a process like *Reverse*, but it is compatible with the fundamental laws. It is also causally explicable. Each glass piece initially starts vibrating because billions of particles in the ground, right where the piece lies, move upwards in (roughly) parallel trajectories and bump into it. Once the glass shards start making small jumps, their motions get reinforced because every time a glass shard touches the ground the molecules in the ground bump into it. The pieces thus keep gaining momentum until they jump up toward the wall-opening. We can, in principle, give a complete description of how the earlier state of glass pieces on the floor causally evolves into the later state of an intact window. But this explanation is practically uninteresting as well as theoretically unsatisfying.

It is practically uninteresting because it needs to appeal to complicated microscopic features. For instance, each glass shard starts moving because of the coordinated trajectories of billions of particles in the floor, which all move parallel and bump into the glass shard at the same time. But whether the particles in the floor are in this kind of motion does not make any macroscopic difference. There is no observable difference in the floor between its state at the time when a piece starts moving and its state a few seconds earlier.

Given our epistemic limitations, we cannot specify the microscopic details in virtue of which the earlier state causes the later intact window. We can say that the particles are coordinated in the kind of way that causes the glass pieces to jump up in the right way. We cannot, however, intrinsically specify the relevant microstate independently of their effect because we do not know enough about microstates. Consequently, we cannot utilize these explanations for prediction or manipulation. We cannot know when the floor is in that precise microstate, except by observing the effect. Moreover, we cannot prepare a system,

such as a floor with glass pieces lying under a wall-opening, to be in a microstate that causes an intact window later.

These practical limitations arise because, as mentioned above, the causal relation between the earlier state of the glass pieces on the floor and the later state of the intact window is extremely sensitive. So the precise microstate matters to why the glass pieces constitute a window, which makes the explanation inaccessible to us. Just as with control, these limitations would not arise for a Laplacian demon, who has full knowledge of the present circumstances and unlimited calculating powers. This demon could specify exactly how each particle needs to be arranged for the effect to happen.

Even a Laplacian demon, however, would find these explanations unsatisfying because a description of how the glass shards on the floor causally evolve into an intact window would lack many objective virtues of good explanations. First, good explanations minimize inexplicable coincidences. But even a full causal description of *Reverse* would contain many features that, in Dummett's words, "cry out for explanation" (Dummett 1964, 339). For instance, why do all the glass pieces start moving at around the same time? And why do they collide exactly in a way that makes them constitute a window? The complete causal description tells us why each glass piece moves exactly as it does at the time it does, but it does not tell us why the motions of the individual pieces are coordinated such that they end up forming a window. No matter how far we trace back the causal story, we would still lack an explanation of why there is such a remarkable coordination between independent

processes, such as the motions of the individual glass pieces.[36] A world in which ordinary processes happen in time-reverse would be full of miraculously-looking coincidences.[37]

A second virtue of good explanations is that they subsume phenomena under robust generalizations. Good explanations do not just specify a causal mechanism; they tell us how one variable would have been different if other variables had been different (cf. Woodward 2003). Now, we can in principle enrich the causal description of *Reverse* to include information about how, for instance, the later state of the window depends on the earlier state of the glass pieces. But because the causal processes are so sensitive, these dependencies would also be extremely sensitive. The smallest change in the background conditions (such as the state of the floor or the motions of the air) and the dependence would no longer hold. So backward-looking causal explanations lack robustness.

Third, some accounts link explanation to unification such that good explanations instantiate patterns that account for multiple phenomena in a uniform manner (cf. Friedman 1974; and Kitcher 1989). In a world where ordinary processes happen backwards, it will be quite common that glass pieces align themselves into windows. However, as seen, a causal explanation of the constitution of the window in terms of the earlier glass pieces has to be tailored to the microscopic details of the earlier state. Therefore, there would be no interesting unifications because these details are different in each case

[36] Several philosophers have observed that explanations of earlier events in terms of later events would be full of inexplicable coincidences. See Dummett (1964); Hausman (1998), chapter 8; and Owens (1992).

[37] Penrose rightly points out that if such cases were widespread, we would be tempted to endorse teleological explanations in terms of a later *telos* (Penrose 1989, 307). Imagine glass shards on the floor would regularly constitute windows at later times and that there would be nothing more we could say except that the particular microstate in each case leads to just this behavior. It would then be tempting to say that the reason why we so often find this conspiratorial behavior of individual glass pieces is for the sake of there being intact windows later.

So descriptions of the kind of causal processes my theory entails in the backward direction lack objective virtues of good explanations. And because *Reverse* it typical in this respect, the account generalizes to other processes in our universe and thus explains why we causally explain later events in terms of earlier events but not *vice versa*.[38]

6 Conclusion

I have argued that the view of causation that best fits with fundamental physics is that causation is bi-directional. According to this view, causation goes in both temporal directions, but it is still time-asymmetric because it has a different character in the forward than in the backward direction. I have shown that this difference in character accounts for the time-asymmetries of control and explanation and hence for why we experience causation as only going in the forward direction. Therefore, bi-directional causation is the most natural view in light of fundamental physics and is also compatible with our experience.

[38] Frisch (2005, 2009) argues that certain good explanatory practices in electromagnetism require causation to be time-asymmetric. His cases are however extremely controversial. See North (2007) and Norton (2009) for criticism of Frisch's argument.

REFERENCES

Albert, D. (2000) *Time and Chance*. Cambridge: Harvard University Press.

Bishop, R. (2009) "Chaos," *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/fall2009/entries/chaos/

Carroll, S. (2010) *From Eternity to Here: the quest for the ultimate theory of time*. New York: Dutton.

Cartwright, N. (1979) "Causal Laws and Effective Strategies," *Nous* **13**: 419-437.

Dowe, P. (2000) *Physical Causation*. Cambridge: Cambridge University Press.

Dummett, M. (1964), "Bringing about the Past," *Philosophical Review* **73**: 338-359.

Earman, J. (1976) "Causation: a matter of life and death," *Journal of Philosophy* **73**: 5-25.

Elga, A. (2000) "Statistical Mechanics and the Asymmetry of Counterfactual Dependence," *Philosophy of Science* **68** (Supplement): 313-324.

Field, H. (2003) "Causation in a Physical World," in: Loux, M. and D. Zimmerman (eds.) *Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 435-460.

Friedman, M. (1974) "Explanation and Scientific Understanding," *Journal of Philosophy* **71**: 5-19.

Frisch, M. (2005) *Inconsistency, Asymmetry, and Non-Locality*. Oxford: Oxford University Press.

Frisch, M. (2009) "'The most Sacred Tenet'? Causal Reasoning in Physics," *British Journal for the Philosophy of Science* **60**: 459-474.

Greene, B. (2005) *The Fabric of the Cosmos*. New York: Alfred A. Knopf.

Hall, N. (2004) "Rescued From the Rubbish Bin: Lewis on Causation," *Philosophy of Science* **71**: 1107-1114.

Hall, N. (2005) "Causation," in: Jackson, F. and M. Smith (eds.) *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press, 505-533.

Hausman (1998) *Causal Asymmetries*. Cambridge: Cambridge University Press.

Healey, R. (1983) “Temporal and Causal Asymmetry,” in: R. Swinburne (ed.) *Space, Time, and Causality*. Dordrecht: D. Reidel Publishing, 79-103.

Horwich, P. (1987) *Asymmetries in Time*. Cambridge: MIT Press.

Ismael, J. (2011) “Temporal Experience,” in: Callender, C. (ed.) *The Oxford Handbook of Philosophy of Time*. Oxford: Oxford University Press. 460-484.

Kitcher, P. (1989) “Explanatory Unification and the Causal Structure of the World,” in: Kitcher, P. and W. Salmon (eds.) *Scientific Explanation*. Minneapolis: University of Minnesota Press, 410-505.

Lange, M. (2002) *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass*. Malden: Blackwell Publishing.

Lewis, D. (1986a) “Counterfactual Dependence and Time’s Arrow,” in: *Philosophical Papers*: *Volume II*. Oxford: Oxford University Press, 32-52.

Lewis, D. (1986b) “Causation,” in: *Philosophical Papers: Volume II*. Oxford: Oxford University Press, 159-213.

Lockwood, M. (2005) *The Labyrinth of Time*. Oxford: Oxford University Press.

Loewer, B. (2007) “Counterfactuals and the Second Law,” in: Price and Corry (2007), 293-326.

Loewer, B. (2012) “Two Accounts of Laws and Time,” *Philosophical Studies* **160**: 115-137.

Maudlin, T. (2007) *The Metaphysics Within Physics*. Oxford: Oxford University Press.

Ney (2009) “Physical Causation and Difference-Making,” *British Journal for the Philosophy of Science* **60**: 737-764.

North, J. (2007) “Book Review: Mathias Frisch, Inconsistency, Asymmetry, and Non-Locality,” in: *Philosophy of Science* **74**: 555-558.

North, J. (2008) “Two Views on Time Reversal,” *Philosophy of Science* **75**: 201-223.

Norton, J. (2007) “Causation as a Folk Science,” in: Price and Corry (2007), 11-44.

Norton, J. (2009) “Is There an Independent Principle of Causality in Physics?,” *British Journal for the Philosophy of Science* **60**: 475-486.

Owens, D. (1992) *Causes and Coincidences*. Cambridge: Cambridge University Press.

Papineau, D. (1985) “Causal Asymmetry,” *British Journal for the Philosophy of Science* **36**: 273-289.

Papineau, D. (1992) “Can We Reduce Causal Direction to Probabilities?,” in: Hull, D. M. Forbes, and K. Okruhlik (eds.) *PSA 1992 vol. 2*. East Lansing: Philosophy of Science Association, 238-52.

Penrose, R. (1989) *The Emperor’s New Mind*. Oxford: Oxford University Press.

Price, H. (1996) *Time’s Arrow and Archimedes’ Point*. Oxford: Oxford University Press.

Price, H. (2007) “Causal Perspectivalism,” in: Price and Corry (2007), 250-292.

Price, H. and R. Corry (2007) *Causation, Physics and the Constitution of Reality*. Oxford: Oxford University Press.

Reichenbach, H. (1956) *The Direction of Time*. Berkley: University of California Press.

Russell, B. (1913) “On the Notion of Cause,” *Proceedings of the Aristotelian Society* **13**: 1-26.

Tooley, M. (1987) *Causation: A Realist Approach*. Oxford: Oxford University Press.

van Fraassen, B. (1993) “Armstrong, Cartwright, and Earman on Laws and Symmetry,” *Philosophy and Phenomenological Research* **53**: 431-444.

Woodward, J. (2003) *Making Things Happen*. Oxford: Oxford University Press.

WHY WE CANNOT CONTROL THE PAST

Chapter Two

1 Introduction

Here are two things I cannot do: I cannot fly and I cannot play Beethoven's Waldstein sonata. The first inability is physical. Given my physical constitution and the laws of nature, none of my decisions would result in me flying. The second inability, however, is not physical but epistemic. It is not physically impossible that I play the sonata. But I still cannot play it because I do not *know how* to play the sonata.[39]

A third thing I cannot do is that I cannot control the past. If I want to spend my next vacation in Paris, there is a lot I can do. I make a hotel reservation, book a flight, etc. But if I want to have spent my last vacation in Paris, there is nothing I can do now. In general, control has a temporal arrow. Our limited control over the future contrasts with a complete lack of control over the past. But why can we not control the past?

Intuitively, my inability to control the past is a physical inability. On our common-sense view, I cannot control the past because none of my decisions would bring about any past outcome. But in this paper I will argue that our inability to control the past is at least partly epistemic, by developing a richer account of what it means for an agent to control an

[39] My distinction between physical and epistemic abilities is similar to Goldman's distinction between epistemic and non-epistemic abilities. See Goldman (1970), 203pp.

outcome. I will show that even if our decisions *did* cause past outcomes, we still could not control the past. Control in the relevant sense requires an unobvious sort of knowledge, and we would lack this knowledge even if our decisions did cause past outcomes.

This result is not just of intrinsic interest. Ordinarily, we take it for granted that causation has an objective temporal direction: causes precede but do not succeed their effects. But many philosophers of physics deny that time-asymmetric causal relations are part of the objective physical world.[40] A main reason against this view is that the direction of causation intuitively underwrites practical asymmetries such as the arrow of control.[41] My account shows that we cannot control the past regardless of whether causation has an objective temporal direction. So our inability to control the past does not account for or against any metaphysical view on the direction of causation. Moreover, my account provides a richer model of what it would take to control the past and why we cannot do it.

In section 2, I show that the kind of control operative in why we cannot control the past is more than just causal influence but also involves a particular sort of knowledge. In section 3, I spell out what it takes for us to have control in this richer sense. In section 4, I set up a thought-experiment where I assume that causation runs backwards and our decisions cause past outcomes. In sections 5 and 6, I show that even under the assumptions of the thought-experiment we still could not control the past because we would lack the necessary knowledge.

[40] Frisch (2012) lists the following philosophers who deny that causation has an objective temporal direction: Suppes (1970), Healey (1983), van Fraassen (1992), Field (2003), Price (1996, 2007), Norton (2003), and Earman (2011).

[41] For instances, Cartwright (1979) argues that objective causal facts are needed to ground the objective distinction between actions that are and are not effective strategies towards a desired end.

2 Two notions of control

The ordinary, causal account of the arrow of control has two parts: a model of control that connects control to causation, and the assumption that causation has an objective temporal directionality.

Causal Account. We cannot control the past, because:

(i) Control consists in our decisions causally influencing an outcome (*Causal Model of Control*); and

(ii) our decisions cause later events but not earlier events (*Directionality of Causation*).

So just as we cannot control, for example, the trajectory of Mars because our decisions do not causally influence Mars's position, we cannot control the past because our decisions do not cause past outcomes.

One way of rejecting the Causal Account is by rejecting the Directionality of Causation. Some philosophers argue that there is no objective physical basis for the perceived time-asymmetry of causation. Our asymmetric causal concept then merely reflects psychological facts about us, but it does not latch on to any objective physical difference relevant to how our decisions affect later and earlier outcomes.[42] So the Directionality of Causation is less certain than we ordinarily think.

[42] See my discussion in chapter 1. In addition, Huw Price in particular has recently argued that the perceived causal asymmetry is merely a feature of our perspective that we project onto the world (cf. Price 1996, 2007). For related views, see Earman (1976) and Healey (1983).

But in this chapter will I reject the Causal Account by rejecting the Causal Model of Control. I will argue that the Causal Account fails because our inability to control the past is independent of whether causation has a direction. Even if our decisions did cause past outcomes, we still would lack control over these outcomes.

According to the Causal Model of Control, control just means causally influencing an outcome. In this sense, the thermostat controls the temperature and the on-switch controls whether the TV is on. If our decisions cause past outcomes, then we have this kind of control over the past.

But the puzzle about our inability to control the past involves a different notion of control. The arrow of control is about agency. We can bring the future (to whatever limited extent) in line with our desires but not the past. If I desire certain future outcomes, I can do things to make the world conform to my desires. But if I desire certain past outcomes, I can do nothing to make the world conform to these desires.

Control in this stronger sense is more than causal influence. In many cases, we can causally influence outcomes but still lack the ability to bring these outcomes in line with our desires. Here is an example:

Hurricane. A platitude from chaos theory says that my current decision to clap my hands can cause a hurricane in Chile six weeks from now (cf. Bishop 2009). An otherwise identical state of the world, except without my clapping, would not lead to a hurricane. But I still cannot control hurricanes in Chile.

My decisions causally influence hurricanes in Chile. Yet, there is a clear sense in which if I want a hurricane to hit Chile six weeks from now, there is nothing I can do to satisfy that desire.

Hurricane shows that the kind of control that allows you to bring outcomes in line with your desires is more than causal influence. It also requires a certain sort of knowledge. Hurricanes are extremely difficult to predict because their occurrence depends on the exact air currents around the globe. My decision to clap influences these air currents. That is, there are circumstances where my clapping conspires with the air currents in the right way to cause a hurricane. But there are also circumstances where the air currents are poised to lead to a hurricane otherwise and my decision to clap would prevent it.

Moreover, I cannot know, at the time of my decision-making, which of these circumstances I am in. So I cannot know whether my decision to clap would make the hurricane objectively more likely than my decision not to clap. Objectively one of my decisions makes the hurricane more likely than the alternatives, but I still cannot do anything to make the outcome conform to my desires because I cannot *know* which of my available decisions that is. So I lack control.

Call cases where my decisions causally influence an outcome but I cannot know which decision would cause the outcome "hurricane cases." Hurricane cases are widespread. Each of our decisions causally influences numerous outcomes without us being able to know how they influence these outcomes. For example, my current decision to move my arm causally influences whether some air molecule in this room will hit a particular gray dot on the wall. Yet, I cannot know whether my decision will cause this outcome.

Control thus has not only a causal but also an epistemic dimension. Without knowledge of which decision would cause an outcome, an agent cannot exploit causal relations to make the world conform to her desires. An adequate model of control has to take into account this knowledge. I will call this richer model of control the "Agent Model."

Agent Model. Control consists in (i) our decisions causally influencing an outcome, and (ii) knowledge that one of our decisions makes the outcome objectively more likely than the alternatives.

The Agent Model explains why we lack control in hurricane cases. We cannot know which decision (if any) makes the outcome objectively more likely than the alternatives. I will argue that our inability to control the past is equally due to our lack of knowledge and thus explained by the Agent Model. That is, past outcomes are analogous to hurricanes in Chile.

According to the Agent Model, control is partly an epistemic ability in the same sense in which the ability to play a certain sonata is epistemic. It is physically possible for my body to go through movements that produce the right sounds. But I still cannot play the sonata because I do not know how to make my body perform these movements. The knowledge I lack does not have to be propositional knowledge or knowledge-that. An agent who can play the sonata might not even in principle be able to articulate this knowledge. The required knowledge fits better what philosophers have called "knowledge-how." To be able to play the sonata an agent has to know *how* to play the sonata. In the same sense, my lack of control over the hurricane is partly due to lack of knowledge-how.

The data about the arrow of control are compatible with our ability to causally influence the past—as long as we cannot know how our decisions influence the past. The data are that we cannot shape the past in accordance with our desires and that we do not experience ourselves as acting on the past.

Both features are shared by hurricane cases, where we do have causal influence. First, as shown, our lack of knowledge prevents us from shaping the weather according to our desires. We could not know which of our decisions would cause rather than prevent the desired weather conditions. Second, we do not experience ourselves as acting on the weather. Our knowledge that we have causal influence is inferential. We know that the weather is due to the earlier air conditions, and we know that our decisions affect the air conditions. But we do not experience ourselves as acting on the weather. For example, we never perceive our decisions as having a direct impact on hurricanes in Chile. So to explain the arrow of control, it is enough to show that past outcomes are analogous to hurricanes in that we cannot know how our decisions influence past outcomes.

3 Agent-control, sensitivity, and knowledge

What about hurricane cases makes it that we lack the knowledge required for agent control? I will argue that a certain ignorance is part of our decision-making, and this ignorance leads to lack of control if causal relations have a certain character.

Our knowledge of the circumstances of our own decisions is limited in two ways. First, we do not know the details of our own decisions. We deliberate about our decisions under macroscopic descriptions. For example, we deliberate whether or not to *throw a stone at a window*. But we are ignorant of the exact details of how these decisions would be

realized. For example, at the moment of decision-making I am ignorant of how, for example, the decision to throw a stone, if I were to make it, would be realized by the neurons in my brain.

Second, we know only macroscopic features of our environment and are thus largely ignorant of the background conditions, that is, what the rest of the world is like at the moment of our decision. For instance, before deciding to throw a stone at a window, I can check that no person or wall blocks its presumed trajectory, that the stone has appropriate weight, and that there is no strong wind. But I will remain ignorant of many aspects of the background conditions, including microscopic features. For instance, I do not know the precise wind conditions, the microscopic composition of the stone, the exact state of my muscles and tendons, or anything much of what is going on in Chile.

So when making decisions, we only know that our decisions and the background conditions fit a certain macroscopic description, but we do not know most of the details. More precisely, when I deliberate between decisions *D* and *D** I do not know which precise microstate would realize a given decision, for example, which of microstates *d1* through *dn* choosing *D* would put me in (and similarly for *D**). Moreover, I also do not know which background conditions *B1* through *Bn* obtain. All I can know is that the relevant states will satisfy some macroscopic description and thus be within a certain margin. But given my epistemic limitations, I cannot 'close' this “margin of ignorance” entirely. These margins of ignorance are illustrated in the figure below.

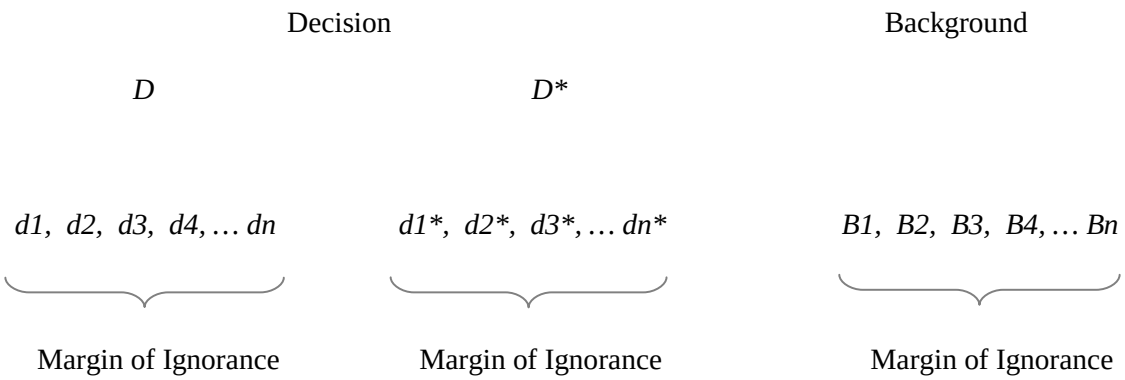

Figure 1 - *Realization of Decisions*

Because of this ignorance, we cannot know the exact consequences of our decisions. A scenario where my decision *D* is realized as *d1* and the background conditions are *B3* causes different effects than a scenario where my decision is realized as *d4* and the background conditions are *B2*. For example, the exact future effects of my decision to throw a stone at a window depend on the details of how the decision is realized, the weight of the stone, the exact air conditions, how the molecules of the stone are arranged, etc. But I cannot know these things when I decide whether to throw a stone.

Due to these margins of ignorance, we can lack control even over outcomes that our decisions causally influence. Imagine you have the choice between decisions *D* and *D** and that given background conditions *B3*, decision *d14* would cause the outcome, but given background conditions *B32*, decision *d2** would cause the outcome, and no other combinations of circumstances would cause the outcome. Your decision thus causally influences the outcome. But to actually control the outcome you would need extremely detailed knowledge of the circumstances of your decision. You would have to know, first, which exact realization of your decision would cause the outcome in which background conditions; second, which background conditions are actual; and, third, whether your

decision *D*, if you made it, would be realized, for example, by *d14*, rather than by some alternative compatible microstate.

In hurricane cases, we lack this knowledge. In some background conditions my decision to clap (if the microscopic details turn out the right way) would cause a hurricane in Chile six weeks later. But I neither know what these background conditions are, nor whether they actually obtain, nor whether my decision, if I made it, would be realized in the right way. The atmospheric conditions all around the planet would have to be in just the right state, and my decision would have to pan out in just the right way, for my clapping to cause the hurricane. But for all I know the circumstances might be such that the hurricane would happen without my clapping and clapping would prevent it. Hence, I lack control.

So we lack the knowledge required for control in hurricane cases because the causal relations are extremely sensitive. I define *sensitivity* and *robustness* of causal relations as follows:

A causal relation from *c* to *d* is ***robust***, just in case: there are many close variations of *c* or the background conditions that would still cause an outcome of the same type as *d*.

A causal relation from *c* to *d* is ***sensitive***, just in case: there are no close variations of *c* and the background conditions, or only very few variations, that would still cause an outcome of the same type as *d*.

Sensitivity and robustness come in degrees. They are a matter of how much variance the cause or the background conditions allow while still resulting in an outcome of the same type.

In hurricane cases, the causal relations between my decision and the respective outcomes are extremely sensitive. For example, suppose my actual decision to clap does cause a hurricane. For this to be the case, the background conditions and the details of my decision have to be orchestrated in the right way. The same decision in slightly different background conditions would no longer cause a hurricane, and neither would a slightly different decision in the same background conditions.

In contrast, the causal relation between your decision to throw a stone at a window and the shattering of the window is much more robust. As long as the background conditions satisfy certain macroscopic features (no obstacles, the window shutters are open, no strong wind, etc.), your decision causes the shattering regardless of how exactly the details of your throw turn out (the exact weight of the stone, the precise angle, etc.). Moreover, any decision that causes the shattering would (most likely) still cause it if the background conditions were slightly different. So the causal relation is very robust.

Given our limited knowledge about the present circumstances of our decisions, whether we can control an outcome depends on how sensitively our decisions cause it. The more sensitive the causal relation, the more knowledge about the circumstances of your decisions is required to know which decision would cause the outcome. For instance, it is relatively easy to learn when a decision would cause a window shattering. In many cases, where the window does shatter, you decided to throw a stone at the window earlier. And in most cases where the window does not shatter, you did not decide to throw a stone earlier.

Moreover, you can easily find out enough about the environment to make even finer distinctions, such as when the window did not shatter despite your decision to throw (for example, when there was an obstacle in the way). In general, when causal relations are robust, you can know whether a situation will result in a given outcome from knowledge of just the earlier macroscopic features.

In the hurricane case, where the causal relation is very sensitive, there is no such macroscopic pattern. Suppose you closely follow the weather reports for Chile, trying to find a salient difference between cases where a hurricane occurred six weeks later and ones where no hurricane occurred. You will find no salient difference at the macroscopic level. Because the causal relation is so sensitive, whether a scenario leads to a hurricane or not depends on the exact microscopic features of the case. But agents like us lack the epistemic capacity to distinguish scenarios based on these features. Hence, you cannot know what decision to make in order to cause the desired outcome.

The more sensitive a causal relation the more knowledge an agent needs to exploit it for controlling the outcome. But, as seen above, there are limits to how much we can know about the circumstances of our decisions. In hurricane cases, the causal relations are so sensitive that control would require more knowledge about our present circumstances than we can have. Control thus breaks down despite causal influence. In the remainder of the paper, I will argue that causal relations in the backward direction are like hurricane cases. Even if our decisions caused past outcomes, these causal relations would be so sensitive that we could not have the knowledge required for control.

But before that, I want to linger a bit on how the Agent Model enriches our understanding of control. The Agent Model explains why some outcomes are more difficult

to control than others. For example, why do we admire artists who can play a particular piece of music with perfection or a golfer who can accurately hit a ball? Dependence in these cases is very sensitive. If the musician moves her fingers just a bit differently, the note will not come off right, and if the golfer swings just slightly differently, the ball will not land close to the hole. Achieving such perfect calibration between one's decisions and arm movements requires a lot of practice and innate skill. The Agent Model of control explains this as the acquisition of knowledge of what decisions lead to the desired outcome.

Moreover, the Agent Model explains how control can come in degrees. The professional golfer has much more control over where the ball lands than I have. This difference is not captured by whether our decisions causally influence the respective outcome. My decision to strike my ball influences where it lands just as much as the golfer's decision to strike her ball influences where it lands. By sheer luck, I might even manage to hit the ball closer to the hole than she. But, due to talent and practice, the golfer knows a lot better than I about which of her decisions leads to the desired outcome. So she can produce the desired outcome more reliably. The Agent Model thus provides a richer understanding of control. Control is not just a matter of whether our decisions causally influence an outcome; it also matters how sensitive the causal relation is and what we can know.

4 Running causation backwards

To demonstrate that our inability to control the past is due to a lack of knowledge, I will assume as a thought-experiment that causation runs backwards, and show that we still could not control the past in such circumstances.

That causation goes only in the forward direction is a substantial metaphysical assumption that some philosophers have denied. Most mature physical theories have time-symmetric laws. According to these laws, what happens earlier lawfully depends on what happens later in the same way in which what happens later depends on what happens earlier. So a different present would lawfully entail not only a different future but also a different past.

Specifically, the mathematical equations stating these laws allow us not only to predict the future states of closed systems (i.e., systems that are, approximately, isolated from their environment) from their present state, but also to retrodict their past states. We can take the state of the world at any one time, where, for instance, a stone hits a window, and run the laws forwards to a later state where the window is shattered. But we can equally take the shattered state of the window and run the laws backwards to an earlier state where the window is intact.

A popular position in the philosophy of physics is the following: While it is natural to assume that the world evolves forwards and earlier events cause later events but not *vice versa*, this time-asymmetry has no basis in the fundamental laws. Moreover, the fundamental physical laws tell us the complete story of how what happens at one time depends on what happens at other times. Consequently, if these laws are time-symmetric, then there is no place for a time-asymmetric dependence relation of the kind causation would intuitively have to be. Our bias for assuming that causation runs forwards is thus merely psychological. Objectively, earlier events depend on later events in the same kind of way in which later

events (such as the shattering of the window) depend on earlier events (such as a stone flying toward the window).[43]

If this view of causation is right, then my thought-experiment requires no counterfactual stipulations. Because it is empirically plausible that the actual laws are time-symmetric, it is, on this view, plausible that there is no causal difference between the forward and the backward direction even in the actual world.

However, I do not have to commit myself to this view of causation. Many philosophers argue that causation runs in the forward direction even if the laws are time-symmetric. The directionality of causation, on these views, is either grounded in other time-asymmetric features of our universe or taken as primitive. Prominent candidates for grounding the time-asymmetry of causation include: primitive dependence (cf. Frisch 2005), causal powers (cf. Cartwright 1989), and an intrinsic direction of time (cf. Maudlin 2007 and Tooley 1997). So causation in the actual world runs forwards because these features 'point' in the forward direction: time runs forwards, causal powers are directed forwards, etc.

But we can still make sense of causation hypothetically running backwards by reversing these features. Take whatever entities you think ground causation in the forward direction and replace them with their time-reverse counterparts. For example, if you think stones flying toward windows cause glass shards on the floor because stones have the causal power to shatter windows (at later times), make it instead that glass shards have the causal powers to constitute windows (at earlier times). Or, if you think causation involves primitive dependence, make it that the earlier state of the intact window primitively depends on the later state of the glass lying on the floor. My thought-experiment thus stipulates a world

[43] See Earman (1967), Price (1996, 2007), and van Fraassen (1993).

where the laws are time-symmetric and all other features that might be relevant to causation point in the backward direction.

In the world of my thought-experiment, our decisions cause past outcomes. For the sake of the argument, I will use the following as a criterion for causation:

Difference-Making. An event *c* at time *t* causes another event *d*, if: an exact copy of the state of the world at time *t*, except without *c*, no longer lawfully entails the occurrence of *d*.

Causes make a difference to their effects: without them, the other events would no longer lawfully entail the effect. For example, my decision to throw a stone causes the shattering of the window just in case an exact copy of the actual state at the time, except without my decision to throw a stone at the window, does not entail the shattering of the window.

It is extremely plausible that difference-making is, given my other stipulations, sufficient for causation in the backward direction. It is, however, not necessary. In cases of redundant causation, the causes do not make a difference to their effects. For instance, assume Billy and Suzy each throw a stone at the same window. Suzy is a bit faster. Her stone hits the window first and causes its shattering. Billy's stone only goes through empty air. Suzy's throw causes the shattering, but it does not make a difference. An exact copy of the actual state at the time of Suzy's throw, except without Suzy's throw, would still lawfully entail the shattering. Billy's stone would hit the window.

So there are more causes than there are difference-makers. However, that my criterion leaves out these potential effects of our decisions on the past does not undermine my

argument that backward causation would give us no control over the past. In fact, causes that are not difference-making are unsuited for control anyway. If the window would still shatter, even if Suzy did not throw, then Suzy cannot control the shattering of the window by deciding to throw or not throw.

Thus, if I can show that my decisions do not give me control over the past despite difference-making, then I have shown that we still could not control the past even if our decisions did cause past outcomes. If the laws are time-symmetric, then our decisions make a difference to the past. A state just like the actual one, except without some decision of mine, no longer lawfully entails certain past outcomes that actually occur. In the following, I will argue that we still could not control the past in these circumstances.

Some theories of causation, however, deny that difference-making is sufficient for causation even given the stipulations of my thought-experiment. According to these theories, the actual direction of causation is due to time-asymmetric patterns among actual events, such as forking, overdetermination, or independence.[44] Reversing these patterns would radically change the actual past and so make the resulting world unsuited as a test-case for whether we could control the actual past if our decisions caused past outcomes.

But I argue that even on this view of causation our lack of a certain kind of knowledge still plays a crucial role in explaining why we cannot control the past. Every account that appeals to causal facts to account for why we cannot control the past needs to explain how control and causation are related. However, such a story is particularly challenging on this understanding of causation because the time-asymmetries that ground causation according to these views appear rather superficial. Lewis (1986), for instance,

[44] See Dowe (2000), Hausman (1998), Lewis (1986), Papineau (1985), Reichenbach (1956). See Weslake (2006) for an overview and criticism of these accounts.

grounds the direction of causation in a particular kind of time-asymmetric counterfactuals. But many philosophers have pointed out that it is not clear how these counterfactual relate to control (cf. Horwich 1987, 171ff; Woodward 2003, 137). What makes Lewis's counterfactuals more relevant to control than other kinds of counterfactuals? Why, for instance, is difference-making in the backward direction (plus the reversal of the other features mentioned), which is also a form of counterfactual dependence, not relevant to control?

Appeal to causation cannot explain why we cannot control the past unless there is a story of what it is about causation that makes it essential to control. Otherwise, it is not clear why we could not control past outcomes in virtue of our decisions bearing some other relation to these outcomes that is not causal, such as lawful determination. So the Causal Model of Control is insufficiently explanatory even on this last understanding of causation. I will argue that we can explain why relations we bear to past outcomes do not give us control over those outcomes in terms of our lack of the required knowledge.[45]

In the following, I will assume that difference-making in the backward direction is causal and show that we still could not control the past. But even if you think that difference-making is not causal, you should still be interested in why bearing this relation to past outcomes does not give us control.

[45] Albert (2000) and Loewer (2007)'s influential account according to which the time-asymmetry of causation is grounded in the same features of the boundary conditions as the thermodynamic asymmetry, for instance, still appeals to our knowledge to explain why we cannot use dependence in the backward direction for control. Loewer's account is in many respects similar to mine.

5 The sensitivity of backward causation

Even if our decisions did cause past outcomes, we still could not control the past because causal relations in the backward direction would be extremely sensitive. This sensitivity has two sources: first, the sensitivity of backward-directed causal processes in our external environment, and, second, our make-up as agents.

For simplicity, I assume that the fundamental laws of nature are the laws of Newtonian Mechanics. Given Newtonian Mechanics, we can specify the complete state of a system at a time by specifying the position and momentum of each particle. Call such a state a *microstate*. We cannot control microstates even in the forward direction. Because any small difference in our decisions or the background conditions would lead to a different microstate, our decisions cause microscopic outcomes only extremely sensitively. But we often can control whether a system is in a certain coarsely individuated macroscopic state in the future, such as whether a window shatters, an egg gets fried, or whether an electron passes through a slit.[46]

We can control these outcomes because the relevant processes are robust in the forward direction. Processes such as a stone shattering a window lead to the macroscopic outcome of a shattered window irrespective of small changes in the initial state. In this section, I will argue that, while the macroscopic processes that lead to such outcomes are robust in the forward direction, they are extremely sensitive in the backward direction. My discussion follows closely Elga (2000).

[46] This distinction between micro- and macrostates is not necessarily connected to size. For instance, an electron passing through a slit counts as a macrostate because it can be realized by many distinct precise microstates that specify the position and momentum of the electron more exactly.

Most macroscopic processes in our universe are thermodynamically irreversible. These processes only ever happen in one temporal direction. For instance, stones flying towards windows often precede glass pieces and a stone lying on the ground, but states of the latter kind never precede states of the former kind as stages of the same process. Similarly, humans grow older but never younger; cigarettes burn to ashes, but ashes never reconstitute cigarettes; ice cubes melt at room temperature, but water puddles never spontaneously form ice cubes; etc. These so-called "irreversible processes" are associated with an increase in entropy: Entropy in our universe never decreases and typically increases toward the future. This asymmetry is characteristic of macroscopic processes. If we see a video recording of a window shattering, we can immediately tell whether the video is played forwards or backwards.

The Newtonian laws are time-symmetric and deterministic. The state of the world at any one time, together with the laws, entails not just a unique future but also a unique past. Say *S1* is the state of the world at a time when a stone flies toward an intact window. *S2* is the state of the world five seconds later when the window is broken and glass pieces and a stone lie on the floor.

The *S1*-to-*S2* process. The stone hits the window, breaking the bonds between the glass molecules. The glass shards and the stone fall to the ground. With every impact on the ground, they create vibrations on the floor and patterns of diverging waves until they come to rest.

We can take *S1* and run the laws forwards for five seconds to get *S2*. But we can also take *S2* and run the laws backwards for five seconds to get *S1*.

This process in the backward direction from *S2* to *S1* is, however, extremely sensitive to small changes to *S2*. If *S2* were just a tiny bit different, it would no longer entail a macroscopic state of a stone flying toward an intact window five seconds earlier. Moreover, this feature generalizes to all thermodynamically irreversible processes, and hence the majority of macroscopic processes in our universe. These processes are all extremely sensitive to small changes in their final condition.

Because it is very hard to picture processes running in reverse, I will use a trick from Elga (2000). The Newtonian laws are time-reversal invariant, which means that whatever can happen forwards can also happens backwards. So instead of looking at the *S2*-to-*S1* process, we can investigate an analogous process that runs in the forward direction. Let *Z2* be the velocity-reverse of *S2*. That means, *Z2* matches *S2* in all respects, except that the velocity of each particle is reversed: Each particle has the exact same position and moves with equal speed, but its direction of movement is rotated by 180 degrees.[47]

Z2 and *S2* are macroscopically indistinguishable. They are both states where glass shards and a stone lie under a broken window. But, because all velocities are reversed, if you take *Z2* and run the laws forwards for five seconds to a state *Z3*, the particles move in exactly the same way as they would move if you take *S2* and run the laws backwards for five seconds to *S1*.

[47] The discussion can be adapted to take account of more complex time-symmetric laws. For instance, in time-symmetric interpretations of quantum mechanics a more complex operation plays the role of velocity reversal in Newtonian Mechanics.

The *Z2*-to-*Z3* process. Particles in the floor bump into the glass shards and the stone in coordinate patterns. The stone and the glass shards begin to vibrate as air molecules around the room form a series of converging waves. Finally, the glass shards and the stone lift up in a single coordinated motion. The glass shards jump toward the wall opening and, just as the stone passes through, collide in the right way to constitute a window. The stone accelerates away from the window.

The *Z2*-to-*Z3* process is exactly identical with regard to the motions of all particles to the process you would get by running the laws from *S2* backwards to *S1*. So if whether the particles in *Z2* move into a state five seconds later where a stone flies away from an intact window is sensitive to small changes in *Z2*, then whether the particles in *S2* move into a state five seconds earlier where a stone flies away from an intact window is also sensitive to small changes in *S2*. I will thus show that the *S2-to-S1* process is sensitive to small changes in *S2* by showing that the *Z2-to-Z3* process is sensitive to small changes in *Z2*.

The *Z2-to-Z3* process is causally explicable. We would be shocked to see glass shards spontaneously form a window, but the process is in accordance with the fundamental laws and we could causally explain it. Think of the process in terms of its microscopic components. We can explain the movements of each particle of the stone and the glass pieces in terms of its earlier interactions with other particles in the floor and the air. For instance, we can explain why each glass piece starts to vibrate in terms of molecules in the floor and the air bumping into it.

The *Z2-to-Z3* process is extremely sensitive to small changes in *Z2* because for it to lead to the outcome of a stone flying away from an intact window, the component processes

have to be amazingly well-coordinated. For instance, each particular glass piece only starts moving because billions of vibrating molecules that all happen to move upwards in a coordinated pattern bump into it. And for it to gain enough momentum to move upwards toward the window opening, once the glass piece starts vibrating, its movement has to be reinforced by the impact from such coordinated molecules in the floor every time it touches the floor. Without such patterns, the glass piece would never start moving and would not gain further momentum once it has started. Moreover, the movements of every glass piece and the stone have to be coordinate with each other to collide in the right way for the glass pieces to eventually form a window as the stone passes through. Not only must the pieces collide in the right way, their molecules also must be in the right state to form chemical bonds.

Because the overall process leads to its outcome only due to this remarkable coordination among its parts, if just one component is a bit off, the whole process would not lead to its actual outcome. For illustration, take a state that differs from *Z2* only in that a few of the vibrating molecules in the floor move differently. Suppose the molecules, just under where one of the glass pieces lies, move with slightly less momentum. Consequently, this glass piece will move a bit differently. Moreover, this difference will spread out to other molecules in the floor, which then hit the other glass shards differently as well, and so all other glass shards will move differently too. This small change is enough to make it that the glass shards will not constitute a window and, most likely, will not jump upwards at all.

Almost any small change to the air molecules, the molecules in the glass shards, the stone, or in the ground can disrupt the complicated patterns that are necessary for the glass pieces to form a window. So the *Z2-to-Z3* process is incredibly sensitive to small changes in

Z2. And because the *S2*-to-*S1* is analogous to the *Z2*-to-*Z3* process with respect to the motions of all particles, the *S2*-to-*S1* is thus equally sensitive to small changes to *S2*.

Some machinery from statistical mechanics will make this point more precise and allow us to generalize it. We can represent the possible states of a system as points in "phase space." Each point represents a precise microstate the system could be in. The evolution of a system then corresponds to a trajectory through phase space, where each point on the trajectory is the state of the system at a time. Macrostates can be realized by numerous precise microstates. Macrostates thus correspond to regions in phase space, where points within the region represent microstates that are indistinguishable with respect to their macroscopic properties. For instance, if we move around some of the particles in *Z2*, the resulting state is macroscopically indistinguishable from the previous one.

Consider the set of microstates that are macroscopically indistinguishable from *Z2*. All of these microstates are such that a stone and glass shards lie under a broken window. These microstates all differ in what future microstates they entail. Some compatible microstates (such as *S2*) are ones where the glass and the stone behave in ways that are thermodynamically normal, for instance where the pieces keep lying on the floor. But some microstates (such as *Z2*) behave thermodynamically abnormal, for instance, ones where the glass pieces jump upwards and form a window.

The 19th century physicist Boltzmann gave an influential explanation of why most systems in our universe behave in ways that are thermodynamically normal. For instance, why do a stone and glass pieces on the floor usually just keep lying there? Why do processes such as the *Z2*-to-*Z3* process never happen?

Boltzmann argues that microstates that behave thermodynamically abnormal, such as *Z2*, take up only a tiny and disjoint subregion (in phase space) among the totality of states compatible with a given macrostate. For instance, take all the microstates that are macroscopically indistinguishable from glass pieces and a stone lying on the floor. Boltzmann's point is that the overall majority of these microstates entail a future evolution where the glass and the stone will just keep lying there. Microstates such as *Z2* are an extreme minority.

The important point is the following:

Microstates that entail a thermodynamically abnormal evolution take up only a tiny and disjoint subregion among the totality of microstates compatible with the relevant macrostate (cf. Elga 2000, 319).

This point entails that the *Z2-to-Z3* process is extremely sensitive to changes in *Z2*.

Z2 is a thermodynamically abnormal microstate. As such, it falls into the tiny and disjoint subregion. All microstates outside this region are thermodynamically normal microstates that lawfully entail a different macrostate from *Z2*. But that means that if we change *Z2* just slightly into a different microstate, this microstate will almost certainly lie outside the subregion of abnormal states and hence behave thermodynamically normal, thus not leading to a later macrostate of a stone flying away from an intact window. The *Z2-to-Z3* process is thus extremely sensitive to small changes in Z*2*.

The same lesson applies to the *S2-to-S1* process. *S2* is abnormal with respect to its evolution in the backward direction. Most surrounding microstates are such that they do not

lawfully entail an earlier macrostate of a stone flying toward an intact window. This lesson generalizes to all thermodynamically irreversible processes. Statistical Mechanics says that macroscopic systems in our universe are extremely sensitive to small changes to their final states. Because microstates that entail the same macroscopic history only take up a small and disjoint region in phase space, it takes only a small change to such a final microstate and the new microstate will lawfully entail a different macroscopic history. So macroscopic processes in our universe are extremely sensitive in the backward direction.

In contrast, macroscopic processes in our universe are robust in the forward direction. For example, the process from *S1* to *S2* (from the stone flying towards the window to the glass pieces and stone lying on the floor) is robust. *S1* is thermodynamically normal: microstates with the same macroscopic future take up the majority of microstates compatible with its macroscopic properties. So any small change to *S1* is overwhelmingly likely to result in a microstate that will lawfully entail the same future macroscopic outcome.

According to my thought-experiment, there are circumstances where my decisions are among the causes of past outcomes, such as a window shattering. However, any such outcome is extremely sensitive to small changes in the later condition. So any causal relation between my decision and an earlier outcome would be extremely sensitive. If the later conditions were just a bit different, the outcome would no longer occur. And we have already seen that we lack control when causal processes are very sensitive. I could never know, prior to my decision, whether it makes the outcome more or less likely.

6 Our make-up as agents

But there is more to the story. We can tell a deeper story of why we could not know the past effects of our decisions by looking at our internal make-up as agents, in particular the role of our decisions.

Our decisions are relatively small and localized events, such as the firing of certain neurons in our brains.[48] Nonetheless, our decisions robustly cause macroscopic events in the future, such as the position of our hands and feet. Because our decisions cause these body movements largely irrespective of what is going on in our environment, we can internally discriminate which of our decisions cause which future body movements. We even individuate our decisions by their effects in the future, such as a *decision to raise my arm*. So I can control the future positions of my arms because I know which of my decisions will cause which arm movement; and I can then control the more distant future because outcomes such as the shattering of a window are robustly caused by my body movements.

This robust 'magnification' of our decisions to the macroscopic level via our bodies is the basis of our control over the future. Without it, we would lack control over the future. To see this, imagine a person in a horrifying state of paralysis where every neural pathway between her brain and the nerves in her body is damaged or blocked. This person still controls her decisions, but her decisions are no longer reliably magnified into body movements.

Her decisions still causally influence the future. The states of her brain cause, for example, the exact positions of photons and air molecules around her head (because different

[48] Or, at least, our decisions are physically realized by such neural events, or their physical correlates are such neural events.

neural states plausibly make a difference to the temperature of her head), which then cause other future differences. Furthermore, as the hurricane example shows, small differences can lead to big differences in the more distant future. So her decisions could still cause macroscopic outcomes in the future. But she could not know which decision would cause a particular future outcome rather than another. So despite this causal influence, the person would be 'trapped in her body', with no control over her environment.

We are in exactly the same situation with respect to the past. Due to our internal wiring, our decisions do not get robustly magnified to the macroscopic level in the past direction.

This magnification happens in the forward direction because our decisions are related to our body movements by distinctive cognitive and physiological processes. Conscious agency (as we use it in control) involves: a state of uncertainty about one's decision, weighing options, deciding in favor of an option, knowledge of what one is going to do, and then the respective action, which we subsequently remember. At the physiological level, these processes are realized, very roughly, as follows: patterns of neuron firings activate action potentials in our spinal cord, which trigger chemical reactions that lead to muscle contractions. We can control the future because these processes are robust with respect to small changes in our brains or the environment. For instance, the exact temperature of your brain, your blood pressure, and the lactic acid in your muscles (typically) do not matter to the effects of your decision on your future body movements.

But these processes are irreversible. Glass pieces and a stone lying on the floor never spontaneously assemble themselves into a state where a stone hits an intact window. And analogously, the robust causal processes by which our decisions cause our future body

movements never happen in reverse temporal order. We never first perform an action, then give up the relevant intention, deliberate about which earlier action to perform, and then become uncertain about our earlier action. So even if there is backward causation, there is no reason to think that there are any robust causal relations between our decisions and *earlier* body movements.

Without these robust connections, we have no more control over the past than a completely paralyzed person has over the future. Our decisions are just small events in our brains, and it would take a lot of knowledge about the details of our decisions and the exact present circumstances to know when they would cause a particular past outcome. However, this is exactly the kind of knowledge that we lack. So we cannot control the past.

This point is independent of the one established in the last section, viz., that the external processes in our environment are sensitive in the backward direction. Because of this asymmetry in our environment, controlling the past is much harder than controlling the future in that it requires a lot more knowledge. But given our internal wiring, we are also a lot better equipped for controlling the future than for controlling the past. Our decisions get robustly magnified to the macroscopic level only in the future direction. So even if there were robust, macroscopic causal processes in our external environment, the causal processes involving our decisions that happen within our skin still would be sensitive in the backward direction. Hence, our decisions still would cause past outcomes only very sensitively.

The sensitivity of external processes in the backward direction and the sensitivity of our internal decision-making processes in the backward direction can each account for why we cannot control the past. But it is unlikely that they are unrelated. Given that processes in the backward direction are so sensitive, it makes evolutionary sense that evolved creatures

like us are more equipped for controlling the future than for controlling the past. We would lack the tools to control the past even if our decisions did cause past outcomes because we would not know which past outcomes our decisions caused.

The unavailability of this knowledge is obscured by the fact that in general we know much more about the past than about the future. But this knowledge is of a particular sort and would not help us control the past. We know the past from records. In treating present events as records of the past, we assume that they have typical causal histories. For instance, glass shards on the floor are typically the remnants of intact windows, and memories are typically caused by the episode they represent.[49] When we draw inferences about the past, we assume that present facts have come into existence in such a typical way. For instance, memories that are induced by hypnosis are not records of the episodes they represent because they have atypical histories.

Could we know from records that some decision of ours, if we made it, would cause a particular past outcome? The problem is that even events which in ordinary circumstances are records of past outcomes no longer count as records in contexts where we make decision in order to control the past. For instance, glass shards under a wall opening are typically a record of an intact window in the past. But if you intentionally distribute glass shards under a wall opening to make it that there was an intact window earlier, you can no longer assume that the glass shards have this typical past history. It is not generally true that glass shards that have been distributed under a wall opening by an agent are the remnants of an earlier

[49] This point is compatible with the earlier point about sensitive dependence. Glass shards need to be in a very specific state to entail an earlier intact window, but, as it turns, out most glass shards in our universe are in such specific states because they all trace back to very particular earlier conditions. Most of them derive from earlier unbroken windows. Why systems in our universe are in such atypical states with regard to their past evolution is the central puzzle in the foundation of thermodynamics. For accessible introductions to this problem see Albert (2000) and Carroll (2009).

intact window in this very wall opening. Your very decision 'screens off' the record from its typical past history.

Records would only help if you could treat your *decision to do something in order to bring about a past outcome* as a record. But such cases would be extremely atypical. Suppose that, typically, whenever I decide to eat chocolate, I have been stressed earlier. My decision to eat chocolate is thus a record of an earlier stressful event.

But is it also true that when I decide to eat chocolate *in order to make it that I was stressed in the past*, it is more likely that I was stressed in the past than otherwise? There is no reason to think so. Even if it is a true generalization that my decision to eat chocolate is typically preceded by a stressful episode, I have no reason to think that this generalization also applies in cases where I eat chocolate in order to make it that a stressful event occurred in the past. The *ceteris paribus* generalization only holds if my decision is typical with respect to its past history. But when I make this decision in order to control the past, the belief-desire structure that precedes the decision is very different from typical cases. Even if my decision to eat chocolate typically indicates that I was stressed earlier, there is no such connection in cases where I decide to eat chocolate because I intend to make it that I was stressed earlier. This case is atypical with regard to its past history.

Now, there might be cases where a decision reliably indicates some past outcome even if the agent makes the decision intentionally in order to make it that this very past outcome has occurred. But we do not know any realistic cases, and if they exist, we have reason to think that they are very rare.[50] The kind of control we would gain from backward

[50] In fact, such cases would be similar to Newcomb's paradox. See Horwich (1987) for an introduction to the paradox and why realistic cases, if there are any at all, would be extremely rare.

causation is thus at most extremely limited. So our lack of the right kind of knowledge explains why we cannot ordinarily control the past.

7 Conclusion

The main upshot of my paper is that agents like us would be unable to control the past even if our decisions did cause past outcomes. Agents like us trying to control the past is like trying to make the number of grains of sand in the desert even by adding truckloads of sand to it. Of course, adding a truck load might change a previously uneven number into an even number. But it might equally keep the number even, or keep it odd, or even change a currently even number into an uneven number; and you have no reason to think that any scenario is more likely than any other. Similarly, a given decision of yours might cause a desired past outcome, but you have no reason to think that it is more likely to cause than to prevent the outcome. Our decisions thus are too blunt a tool for such a delicate task as controlling the past. Agents with more sophisticated knowledge, such as God, could control the past if causation runs backwards—but we cannot.

I have given an explanation of why we cannot control the past that does not rest on contentious assumptions about the metaphysics of causation. In doing so, I have developed a richer account of control, where control is more than causal influence. The kind of control relevant to agency requires that the agent has a certain kind of knowledge of the effects of her decisions. I have shown that even if our decisions did cause past outcomes, we would lack the knowledge required to use this causal influence for controlling the past.

My main goal here was a deeper account of control, but my argument has also consequences for the metaphysics of causation. Our main apparent evidence that causation is

time-asymmetric is our inability to bring the past in line with our desires. But I have shown that this inability is due to us lacking a certain kind of knowledge and thus compatible with causation running backwards. So our inability to control the past is not evidence that our decisions do not cause past outcomes.

REFERENCES

Albert, D. (2000) *Time and Chance*. Cambridge: Harvard University Press.

Bishop, R. "Chaos", *The Stanford Encyclopedia of Philosophy* (Fall 2009 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2009/entries/chaos/>.

Carroll, S. (2009) *From Eternity to Here: the quest for the ultimate theory of time*. New York: Dutton.

Cartwright, N. (1979) "Causal Laws and Effective Strategies," *Nous* **13**: 419-437.

Cartwright, N. (1989) *Nature's Capacities and their Measurement*. Oxford: Clarendon Press.

Dowe, P. (2000) *Physical Causation*. Cambridge: Cambridge University Press.

Earman, J. (1976) "Causation: A matter of life and death," *Journal of Philosophy* **73**: 5-25.

Earman, J. (2011) "Sharpening the Electromagnetic Arrow(s) of Time," in: Callender (ed.) *Oxford Handbook of Philosophy of Time*. Oxford: Oxford University Press, 485-528.

Elga, A. (2000) "Statistical Mechanics and the Asymmetry of Counterfactual Dependence," *Philosophy of Science* **68** (Supplement): 313-324.

Field, H. (2003) "Causation in a Physical World," in: Loux and Zimmerman (eds.) *Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 435-60.

Frisch, M. (2005) *Inconsistency, Asymmetry, and Non-Locality*. Oxford: Oxford University Press.

Frisch, M. (2012) "No Place for Causes? Causal Skepticism in Physics," *European Journal for Philosophy of Science* **2**: 313-336

Greene, B. (2006) *The Fabric of the Cosmos*. New York: Alfred A. Knopf.

Goldman, A. (1970) *A Theory of Human Action*. Englewood Cliffs: Prentice-Hall.

Hausman (1998) *Causal Asymmetries*. Cambridge: Cambridge University Press.

Healey, R. (1983) "Temporal and Causal Asymmetry," in: R. Swinburne (ed.) *Space, Time, and Causality*. Dordrecht: D. Reidel Publishing Co.

Horwich, P. (1987) *Asymmetries in Time*. Cambridge: MIT Press.

Lewis, D. (1986) "Counterfactual Dependence and Time's Arrow," in: *Philosophical Papers*: *Volume II*. Oxford: Oxford University Press, 32-52.

Loewer, B. (2007) “Counterfactuals and the Second Law,” in: Price and Corry (2007), 293-326.

Maudlin, T. (2007) *The Metaphysics Within Physics*. Oxford: Oxford University Press.

Norton, J. (2003) “Causation as a Folk Science,” in: Price and Corry (2007), 11-44.

Papineau, D. (1985) “Causal Asymmetry,” *British Journal for the Philosophy of Science* **36**: 273–289.

Price, H. (1996) *Time’s Arrow and Archimedes’ Point*. Oxford: Oxford University Press.

Price, H. and R. Corry (2007) *Causation, Physics and the Constitution of Reality*. Oxford: Oxford University Press.

Reichenbach, H. (1956) *The Direction of Time*. Berkley: University of California Press.

Russell, B. (1913) “On the Notion of Cause,” *Proceedings of the Aristotelian Society* **13**: 1-26.

Tooley, M. (1997) *Time, Tense, and Causation.* Oxford: Clarendon Press.

van Fraassen, B. (1993) “Armstrong, Cartwright, and Earman on Laws and Symmetry,” *Philosophy and Phenomenological Research* **53**: 431-444.

Weslake, B. (2006) “Common Causes and The Direction of Causation,” *Minds and Machines* **16**: 239-257.

Woodward, J. (2003) *Making Things Happen*. Oxford: Oxford University Press.

CAUSATION, PHYSICS, AND FIT

Chapter Three

1 Introduction

Causation is central to how we make sense of the world. Not only do we explain events by discovering causal connections, but as Cartwright (1979) points out, to survive we need to be able to distinguish effective from ineffective strategies. And whether something is an effective strategy toward an outcome depends on whether it causes the outcome. For instance, taking a drug is an effective strategy for recovery from a disease only if it causes recovery.

Yet, as Russell (1913) observes, our ordinary causal assumptions seem to fit poorly with how our best physics describes the world.[51] We think of causation as a *time-asymmetric* determination relation between relatively *localized* things. That means we typically assume that causes precede their effects and also that effects are caused by a relatively small set of conditions. But Russell argues that fundamental physics describes the world in terms of lawful determination between very global states and, moreover, makes no distinction

51 I got the useful notions of "fit" and "poor fit" from Schaffer (2010).

between the way in which the past determines the future and the way in which the future determines the past (cf. Field 2003).[52]

Russell's own take on this poor fit is that "the law of causality […], like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." (Russell 1913, 1) This advice to get rid of causation, however, is not feasible because science would be crippled and our survival chances diminished without a causal concept. Even those philosophers who agree with Russell about the limited role of causation in fundamental physics admit that causation is central to understanding and manipulating the physical world.

Still, Russell's observation poses a challenge for theories of causation. The challenge is to explain *why* causation works so well. If there is this mismatch between causation and how fundamental physics describes the world, then why is it a good idea for us to represent the world causally the way we do? Why does causation help us understand and manipulate the physical world, given that it seems to be grossly misrepresenting its structure? My goal in this paper is to answer the question of why it is a good idea for us to have the causal concept we do have given our interests in understanding and manipulating the world.

In particular, I will address this question within the framework of causal models. Recent work on causal modeling has been central to understanding how we can draw causal inferences from statistical data and thus to how causation enables us to understand the physical world. The causal models that scientists use, however, display the same locality and temporal directionality as does our everyday causal reasoning. So I will show why our causal

52 Many philosophers, especially in the philosophy of physics, take this poor fit between causation and fundamental physics to show that there are no objective causal facts. See Norton (2007), Price (1996, 2007), and van Fraassen (1993).

reasoning works so well both in everyday life and in science by showing why it is a good idea for us to build these kinds of causal models.

There is a seemingly obvious answer to this question that appeals to evolutionary pressure. If our ancestors had used a causal concept according to which, for instance, later events cause earlier events but not *vice versa*, they simply would not have survived. So we have an explanation of why we have the causal concept we do have. But this answer still leaves open which physical features of our world create this survival pressure. In this paper, I want to identify which physical features of our world make it a good idea for creatures who want to understand and manipulate this world to have a local, time-asymmetric causal concept. Besides its intrinsic interest such a story would also give us a better understanding of how the special sciences relate to fundamental physics. After all, the special sciences allegedly use causal models to explain patterns in the very same world that fundamental physics describes.

There has been some work on why causal models can be local (Eagle 2007; Elga 2007; and Woodward 2007).[53] But causal modelers typically take the direction of causation for granted. The canonical texts either contain no discussion of the temporal direction of causation (Woodward 2003) or merely allude at further work to be done (cf. Pearl 2000, 59). I will show that locality and temporal directionality are closely related. The same physical features of the world that make it possible for us to build local causal models also explain why it is a good idea for us to build time-asymmetric causal models.

[53] I use "locality" there in the way Field (2003) and Elga (2007) use the term to mean that an outcome (or its probability) is determined by what is happening in a relatively small region of space. This sense of locality is different, though related, to what philosophers of physics have in mind when they talk of non-local interactions and non-local laws.

My account disagrees, on the one hand, with philosophers who argue that our causal thinking has no objective basis (Healey 1983; Price 1996, 2007), and, on the other hand, with philosophers who think that causation is a non-explicable metaphysical primitive (Cartwright 1979; Frisch 2005, 2010; Maudlin 2007; Tooley 1987). There is a long tradition of work on how causation fits into the physical world, starting with Reichenbach (1956). But most of these theories merely show which aspects of fundamental physics our causal concept latches on to, thus analyzing our concept (e.g., Dowe 2000; Lewis 1986a), and leave out the normative question of why it is a good idea for us to have this concept rather than a different concept.[54] Other accounts focus on why our causal concept is useful given our interests in manipulating the world but are not straightforwardly applicable to causal models and our interests in explanation (Albert 2000).[55] My own account not only differs from these theories in detail, but it also explains why it is a good idea for us to have the causal concept we do have and its role in causal explanation by bringing in the framework of causal models.

In section 2, I introduce the framework of causal models. In section 3, I explain Russell's challenge that our causal concept fits poorly with fundamental physics. In section 4, I explain why our local causal models are successful. In sections 5 and 6, I argue for why it makes sense for us to build time-asymmetric causal models.

[54] For example, Horwich (1986) and Woodward (2003) criticize Lewis (1986a) on the grounds that he does not have a plausible account of why we have the kind of causal concept we do have.

[55] But see Hausman (1998) for a theory that could be applied to causal models. There are some close connections between Hausman's proposal and my own proposal. See FN 73.

2 Causal models and the physical world

Causal models take variables (or changes in variables) as the causal relata, where variables represent the different states some part of the world can assume.[56] These models represent causal relations between variables in terms of the technical notion of an intervention. The central idea is that a causal relation between two variables is more than just a correlation, and the difference is that if *X* causes *Y*, then there is a possible intervention on *X* that also changes *Y*; but if *X* and *Y* are merely correlated, then there is no intervention on *X* that also changes *Y*. For instance, the storm and the barometer reading are merely correlated because there is no possible intervention on the barometer reading that would change the storm. In contrast, the earlier air pressure causes the later storm because there is a possible intervention changing the air pressure that would also change the storm.

Structurally, causal models consist of a set of variables and a set of equations that specify how variables change given interventions on other variables. A variable *X* causes another variable *Y* iff there is at least one possible intervention on *X* that would change the value of *Y*. An intervention on *X* with respect to *Y* is a change in *X* that does not cause *Y* (or is in any other way correlated with *Y*) except, if at all, via a causal route that goes through *X*.[57] Thus, a process that changes *X* by changing a common cause of both *X* and *Y* is not an intervention on *X* with respect to *Y*. For instance, changing the barometer reading by messing with its scale counts as an intervention on the barometer reading with respect to the storm; but changing the barometer reading by changing the earlier air pressure is not an intervention

[56] I will use "variables" sometimes for things in the world that stand in causal relations and sometimes for representations of such things. It will be clear in each case what I mean.

[57] Interventions are related to Lewis's notion of a "miracle" (cf. Lewis 1986a). However, there are important differences both in detail and in motivation. See Woodward (2003) for discussion of how the two notions relate.

because it changes the storm via a causal route that does not involve the barometer reading as an intermediary.[58]

Building causal models is of central interest to us for at least two reasons. First, causal models can guide our actions because they encode information about which variables are such that their manipulation is, in principle, a means for manipulating other variables. Interventions allow us to compare a system to a copy of itself which has the same initial condition, except that some variable has been changed. For instance, we can compare the actual meteorological conditions to a system that has the same initial state, except that the air pressure in a certain region is different. This comparison shows how the air pressure affects variables at other times, which tells us which other variables could, in principle, be manipulated by manipulating the air pressure.

Second, causal models can underwrite explanations. We can fit statistical data into causal models and thus explain why certain variables had to have the values they have given the values of earlier variables (or why these values have the probabilities they had). Moreover, we can show how their values would have been different if these earlier values had been different. These kinds of explanations are central to sciences like epidemiology or biology.

The causal models that we ordinarily use have two striking features. First, our causal models are extremely *local*. The variables that we use in typical causal models only represent a small portion of the world at any time. For instance, if we try to model the breaking of a glass of water in Chapel Hill, we will typically assume variables such as the hardness of the

[58] Because interventions are defined in terms of causal constraints, the causal modeling framework is not reductive. Rather than reduce causation, the framework aims to clarify the relationships between different kinds of causal relations and their connection to other relations, such as probabilities and counterfactual dependence. For accessible introductions to causal models see Hitchcock (2009) and Woodward (2007).

floor, the distance it falls, or how thick the glass is. But we will typically not include variables about what is happening in outer space or in Russia at the time.

Second, our models are time-asymmetric. The structural equations that specify how variables depend on interventions on other variables describe how the values of later variables depend on interventions on earlier variables but not *vice versa*. Causal models thus represent the final conditions of a system as dependent on its initial conditions but not the other way around. For instance, we represent the current state of glass as dependent on earlier variables but not as dependent on later variables. Causal models thus reflect our ordinary judgments that causes precede their effects and that only a small number of the events preceding an effect are among its causes.

3 Locality, directionality, and fit

But why are causal models that have these features (and our causal reasoning in general) so successful? It is extremely natural to think that causal models are successful because they latch on to the world's lawful structure. The fundamental physical laws tell us how the state of the world at one time depends on its states at other times. A natural suggestion is that our causal models represent this dependence in a piece-wise fashion by telling us which aspects of some earlier state matter to some aspect of a later state (cf. Hall 2004). Most philosophers therefore think that causation reduces to patterns of lawful dependences in some way or other. This reduction is typically cashed out in terms of minimal

sufficiency, probability raising, counterfactual dependence, or lawful regularities, where these relations are in turn understood in terms of lawful entailment.[59]

But the apparent lesson from Russell (1913) is that due to their locality and time-asymmetry our causal models do an exceedingly poor job at latching on to the world's lawful structure. The mismatch is most striking with regard to time-asymmetry. Most candidates for the fundamental physical laws are deterministic in both temporal directions and so lawful entailment goes equally in both temporal directions. According to these laws, the state of the universe any one time uniquely fixes the state of the universe at all later and earlier times. As a matter of physical law, otherwise identical systems that differ in their final condition also differ in their initial condition, just as systems that differ in their initial condition differ in their final condition (or at least in what probability distribution they assign to the final condition).[60]

We can compare a system to a copy of itself whose initial condition differs with respect to some variable and use the laws to determine how the system would differ from the original system at later times. But we can equally compare the system to a copy of itself whose *final* condition differs with respect to some variable and then determine how the change in this variable affects earlier states of the system. This determination is possible because the fundamental physical laws allow retrodiction just as much as prediction. We can thus use the laws of nature to retrodict what the initial condition of such a system must be like given its final condition.

---

[59] See Hall (2005) for a nice discussion and overview.

[60] I assume that the laws are deterministic, but the problem does not significantly change if we move to probabilistic laws, as long as these laws have the same probabilistic character in either temporal direction. This is the case if the state of the world at any one time fixes a unique probability distribution over all earlier and later states (cf. Field 2003). In this case, systems that differ only in their final conditions also differ (at least typically) in what probability distribution the laws assign over the system's initial conditions.

For example, take a ball that is moving across a billiard table. The fundamental physical laws allow us to predict the final condition of the ball from its initial condition; but they equally allow us to retrodict the initial condition of the ball from its final condition. We can assume the ball's momentum changed at an earlier time and use the physical laws to calculate its new momentum at a later time. But we can equally assume the ball's momentum to be changed at some later time, thus creating a new final condition, and then use the physical laws on this final condition to calculate its corresponding momentum at earlier times. This procedure informs us about how the earlier momentum of the ball depends on its later momentum. Fundamental physics thus allows us to build backward-looking models that systematically tell us how earlier variables depend on later variables.

One might object that these backward-looking models would not be *causal* models. After all, causal models tell us how variables depend on *interventions* on other variables, and interventions are defined as forward-directed causal processes. An important feature of how interventions are typically understood is that they 'break' the lawful connections between later variables and earlier variables such that earlier variables do not depend on intervention on later variables.

But this objection misses the real challenge. The point is not that backward-looking models would be *causal* models. The point is that backward-looking models represent the lawful structure of our universe just as well as our causal models. What needs to be explained, what is called into question, is why it is a good idea for us to adopt a time-asymmetric notion of an intervention in the first place.

In particular, we can easily stipulate a time-neutral notion of an intervention. In such *pure interventions* (as I will call them), we simply assume the value of some variable to be

different without assuming any kind of process (causal or not) that brings about this change. Such interventions are completely time-neutral. For example, in a pure intervention on a billiard ball's momentum we assume its momentum at some time *t* to be different, but hold all other variables at *t* fixed, and then use the laws of nature to propagate the change in this value forwards and backwards to see how variables at other times change in response to this change at *t*.

Pure intervention can play the same role in the forward direction as conventional, time-asymmetric interventions. The causal constraints on interventions are meant to ensure that, if a change in *X* is an intervention on *X* with respect to *Y,* the process that changes *X* does not change *Y* in any other way except by changing *X*. In a pure intervention on *X* there is not any process that changes X; we simply assume *X* is changed, while holding everything else at the time of *X* fixed. Any change in *Y* therefore must be due to the change in *X* because everything else at the time of *X* has been held fixed.

Models based on pure interventions thus make the same predictions in the forward direction as our ordinary, time-asymmetric causal models, but they additionally reflect lawful dependence in the backward direction. On the face of it, our time-asymmetric causal models therefore seem at least misleading and possibly highly inaccurate. They seem inaccurate because they represent only part of the physical dependence there is, and they seem misleading because they suggests that this is all the dependence there is. The challenge is thus why, despite the availability of these time-symmetric models, it is a good idea to adopt time-asymmetric models that only reflect lawful dependence in the forward direction.

Directionality challenge. Why does it make sense for us to prefer time-asymmetric causal models over models that also reflect lawful dependence in the backward direction?

The second mismatch that Russell stresses concerns locality. As Field points out, fundamental physics lacks the extreme locality of our causal models because "no reasonable laws of physics, whether deterministic or indeterministic, will make the probability of what happens at a time depend on only finitely many localized antecedent states." (Field 2003, 439) So just as there is a mismatch between the directionality of causation and the bi-directionality of the laws, there is also a mismatch between the local character of causation and the global character of the laws.

The worry is that, by explaining outcomes in terms of a finite number of relatively localized variables, our causal models leave out important aspects of the physical determination of these outcomes. Suppose a glass falls off a table and breaks. A typical causal model depicts the shattering of the glass as dependent on the glass falling from the table. But whether the glass shatters equally depends on what happens in many other spatial regions not represented in this model. For example, if a large number of air molecules all hit the bottom of the glass shortly before it touches the ground; if the earth's gravitational force were drastically smaller; or, if some burst of energy from outer space pulverized the glass in midair, there would be no glass pieces on the ground even if the glass falls from the table. But our causal models typically do not include such variables.

So our local causal models seem inaccurate from the perspective of fundamental physics because they include only a subset of the variables that are relevant to the physical determination of outcomes; and they seem misleading in that they represent variables as only

dependent on a small number of variables. So why are our local models successful despite leaving out numerous variables that also determine features of the modeled system?

Locality challenge. Why are causal models that include only a small number of relatively localized variables successful?

The locality and the directionality challenge both say that our causal models misrepresent the underlying physical reality because they leave out some of its determination structure. The directionality worry says that our causal models leave out determination in the backward direction; the locality worry says that our causal models leave out determination from happenings in regions not represented by the variables in the model. For these reasons, Russell argues that there is a poor fit between causation and the physical world and we should therefore give up our causal concept.

A response to these challenges needs to show why our actual practice, where we build *local* and *time-asymmetric* causal models, is advantageous for explaining and manipulating the physical world in light of its alleged mismatch with fundamental physics. Responses tend to fall into three groups.

First, Primitivists argue that the poor fit between our causal models and the structure of the fundamental laws does not show a deficiency in our causal assumptions but shows that we have to enrich our understanding of the physical world. Primitivists thus posit primitive causal facts that underwrite our local, time-asymmetric assumptions. Causal models on this view do track these primitive causal facts rather than lawful dependence.[61]

[61] Cartwright (1979), Frisch (2005, 2010), Maudlin (2007), and Tooley (1987) each develop versions of causal primitivism. They differ in how exactly this causal structure relates to our physical theories.

Second, Pragmatists also accept that there is a poor fit between our causal models and fundamental physics. But they argue that representing the objective world is not the point of causal discourse. Causation is an anthropocentric notion that helps agents like us get around in the world and it can fulfill this purpose even if it does not latch on to objective physical facts in any direct way.[62]

Third, Reductionists argue that the poor fit between causation and fundamental physics is only apparent. Our time-asymmetric, local causal concept, *pace* Russell, does latch on to objective physical structures, and we can see how by bringing in statistical features from the boundary conditions rather than just the dynamical laws. Our causal models thus track important physical features of the world.

It is possible to favor different replies to the directionality and locality challenge. For example, someone might think that underwriting the time-asymmetry of our models requires metaphysically primitive, time-asymmetric facts and yet think that the success of local models can be explained reductively or has merely pragmatic reasons.[63] In fact, most philosophers have mainly focused on the directionality challenge. I will argue, however, that there is an important connection between the two challenges. Our local models are successful because there are privileged patterns of lawful dependence between certain local variables. This response turns out also to resolve the directionality challenge because such privileged patterns of lawful dependence obtain in the forward direction but not in the backward

[62] Price (1996, 2007) is the main proponent of this view. See also Healey (1983).

[63] For example, Maudlin (2007) argues that the time-asymmetry of causation is underwritten by a primitive directionality of time. This posit allows him to reply to the directionality challenge without also resolving the locality challenge.

direction. Hence, the availability of these patterns vindicates both locality and directionality.[64]

My view thus contrasts with Primitivism because I argue that we can account for the success of our causal models in light of fundamental physics without positing primitive causal facts. Though some of the assumptions I will make have a pragmatist flavor, I think that my view is still reductive. However, I will not defend this point here, as my goal is to give a physical underpinning of the success of our causal reasoning. Moreover, the distinction between Pragmatism and Reductionism is less clear than it seems. After all, Pragmatists agree that there are objective physical facts that explain why our causal concept is useful; it is just that our concept does not directly represent these objective facts.

4 Locality and invariance

I will start with the locality challenge, specifically why local causal models are explanatory. Why do our local models explain outcomes, as we think they do, despite leaving out multiple variables that matter to the probability of these outcomes?

The answer turns on what makes a model explanatory. I do not have the full story, but at least some good-making features of causal explanations are widely accepted. Many philosophers have argued that an essential feature of successful explanations is a certain modal robustness. To be explanatory, a model has to apply not only in the actual circumstances; it also has to tell us what would have happened in a range of other

[64] Causal models concern type-level causation. In another paper, I argue that causation between token events goes in both temporal directions and the time-asymmetries that account for why only forward-looking causal models are explanatory are merely gradual. See chapter 2 of this dissertation.

circumstances. I will follow a particular development of this idea in Woodward (2003, 2007).[65]

Woodward uses the example of Hook's law, which describes the behavior of springs:

(H) $F = -k X$

X is the amount by which the spring is displaced from its relaxed position; *F* is the force by which the spring resists this displacement; and *k* characterizes the "stiffness" of the spring.

Woodward argues that Hook's law explains why a spring resists its displacement with a certain force because it tells us not just the actual value of the force but also its value in a range of counterfactual circumstances. This modal robustness is defined in terms of interventions and Woodward distinguishes two different aspects of it: *invariance* and *stability*.[66] Invariance concerns interventions on variables that are part of the model; stability concerns interventions on variables that are part of the background conditions and are not represented in the model. For example, the displacement of the spring, *X*, is part of the model based on (H), whereas its temperature is part of the background conditions.

Robustness comes in degrees. (H) is not invariant or stable under all interventions. For example, if *X* is increased too much, the spring gets overextended and the restoring force is significantly less than (H) predicts. Furthermore, (H) also breaks down, for example, given interventions that heat up the spring to an extreme degree. Therefore, (H) is neither

[65] Lange (2000, 2009) also emphasizes modal robustness as a central virtue of explanations. See Reutlinger, Schurz, Hüttemann (2011, Section 6) for a comparison between Lange's and Woodward's account.

[66] Woodward thinks of interventions as forward-directed causal processes in the sense described in section 2. For present purposes, it does not matter whether we think of intervention in this time-asymmetric or in my time-neutral sense.

completely invariant nor completely stable, but it stable and invariant to a high degree. It correctly predicts the resulting force for a wide range of interventions on both the extension of the spring and background conditions such as the exact temperature or location of the spring.

Woodward argues that for a generalization to be explanatory, it is enough that it is invariant and stable for a wide range of interventions.[67] There are at least three deep connections between robustness and explanation. First, in finding robust generalizations we *isolate* explanatory factors and make an outcome appear less coincidental. Suppose the value of a variable *Y* depends on interventions on another variable *X*, but this dependence holds only for very few interventions and in very special background conditions. It then seems that there is no deep connection between the two variables; it seems that the background conditions contribute just as much to why *Y* has its actual value as does *X*. But suppose the dependence is robust such that the value of *Y* depends on the value of *X* in a wide range of interventions. This robustness indicates a deep connection between the two variables that makes the particular value of *Y* less coincidental because the value of *X* determines the value of *Y* in a wide range of circumstances. So it seems less of a coincidence that *X* determines this particular value of *Y* in the actual circumstances.

Second, invariance is important for *manipulation*. If you want to change one variable by changing another variable, then a generalization can only guide you if it still holds given your intervention. Moreover, the generalization is more useful if it holds in a wide range of circumstances because we will often not have full control over the circumstances when we act or not know what they are. Suppose (H) were only true for springs at some exact

[67] There are further constraints, for example that the range of interventions under which a generalization is stable has no gaps. These additional requirements will not matter here.

temperature. It would then be much less action-guiding because it would only tell us how far to displace a spring to create a certain resisting force if the spring has that particular temperature, and we might not know when that is.

Third, robustness matters to the *scope* of a generalization. For example, different springs differ in details, such as their location or temperature. But because Hooke's law is stable, it is approximately true in a wide range of circumstances. We can apply it to many different springs regardless of these differences and gain a unifying account of the behavior of different springs. This connection is important because unification is often considered an important goal of explanations (cf. Friedman 1974; and Kitcher 1989).

These connections show that robustness is a central good-making feature of explanations. Therefore, if I can show that the dependencies between variables encoded in our local causal models are robust, I have given a plausible reason for why these models are explanatory. I thus need to show that local causal models have this robustness. This robustness is surprising because, in accordance with the fundamental physical laws, the probability of an outcome depends on an extremely large number of earlier variables.[68] Why then do the fundamental physical laws allow for local models that are robust?

Before I can show why local, robust models are widely available, I need to make explicit two additional features of causal explanations. First, as mentioned, to be explanatory, models do not have to be robust under all interventions, but only under interventions within a certain range. In addition, Woodward emphasizes that interventions that produce circumstances that are likely or typical are more important than interventions that lead to

[68] For example, even if the speed of light is the upper-bound on how fast signals can travel, this upper-bound would still allow that events anywhere on earth can influence how, say, a glass behaves over the next 0.05 seconds (cf. Elga 2007).

more far-fetched circumstances. In his own words: “In the case of macroscopic causal generalizations we are particularly interested in invariance and stability under changes that are not too infrequent or unlikely to occur, around here, right now, and less interested in what would happen under changes that are extremely unlikely or which seem ‘farfetched’.” (Woodward 2007, 79)

Second, causal explanations in the special sciences and in everyday life involve fairly coarse-grained variables (cf. Field 2003 and Woodward 2007). For example, we might represent a glass of water with a variable whose two values are *broken* and *unbroken*. These values are coarse-grained in the sense that they can be realized in multiple ways by precise microstates. Therefore, typical causal explanations address why systems have certain coarse-grained macroscopic, rather than exact microscopic, properties.

Both features can be justified pragmatically. Woodward points out that circumstances that are likely or typical have a special relevance to our interests in manipulation. If you want to change a variable Y by changing another variable X, you only need to be concerned with whether Y depends on X in circumstances that are likely to obtain around here and now, i.e., at the time and place at which you are acting (cf. Woodward 2007, 80). The same holds for prediction. Similarly, our epistemic capacities only allow us to reliably discern the values of coarse-grained variables. Fine-grained variables indicating the exact microstate of a system would be of no use to us because we could not know which of their values obtain.

However, we can also defend these assumptions on more metaphysical grounds. Coarse-graining is plausible because the special sciences study features of systems that are multiply realizable and are to some degree independent of the system's exact physical realization, such as monetary exchanges. Furthermore, the special sciences describe systems

that are situated in particular environments, such as markets or populations. So, we should focus on the properties they manifest within these environments. For example, the cup on my desk falls down when dropped. It has this feature only as long as the earth's gravitational field does not drastically decrease. Such circumstance, however, would be extremely far-fetched relative to the environment within which I am trying to understand the cup. Therefore, it makes sense that explanations concern coarse-grained variables in typical or likely circumstances.[69]

The relevance of these two features is that in building explanatory causal models we can leave out all variables that either (i) do not change in typical or likely circumstances, or that (ii) do not affect the value of coarse-grained variables. We do not have to include variables in our models that are either almost always constant or whose impact on other variables is so small that they typically do not make a difference to their macroscopic properties. If I can show that most variables in our universe satisfy either (i) or (ii), I have shown why local causal models that are explanatory are widely available.

Elga (2007) provides the required argument. Elga's argument has two parts. First, he argues that, although systems in our universe are subject to a multitude of forces, most of these forces are either "negligibly tiny or nearly constant" (Elga 2007, 109).

Negligibility of distant forces. Distant forces are either extremely small or almost constant.

[69] Which circumstances count as likely or typical is a difficult question. There are two ways to go about it. First, some philosophers argue that we get a global probability distribution over all macrostates from the foundation of statistical mechanics (cf. Albert 2000, chapter 5). We could then use this probability distribution to determine which variables are likely to change. Second, some philosophers argue that we should invoke pragmatic factors to determine which values of a variable count as the default and therefore as to be expected (cf. Hitchcock 2007). Either route could be adopted for present purposes.

There are four kinds of forces in our universe: strong, weak, electromagnetic, and gravitational forces. Strong and weak forces are negligibly small at distances greater than the subatomic level. Electromagnetic and gravitational forces, in contrast, can be strong even at great distance. However, electromagnetic forces are still weak around here because there are no massively charged bodies. Gravity is strong, but it is almost constant and does not change drastically at nearby locations or over time.

Second, Elga argues that extremely small forces are typically irrelevant to the future behavior of macroscopic objects. For instance, if a glass falls from a table, then small differences in its microstate are very unlikely to make a difference to whether it will break or not. Call this fact “macroscopic stability.”

Macroscopic stability. Small microscopic differences are extremely unlikely to affect the future macroscopic behavior of a macroscopic system.

Because small forces typically do not affect the macroscopic behavior of systems, they also do not affect the values of coarse-grained variables that describe macroscopic properties of a system, such as whether a glass is broken or unbroken.

Elga's argument shows that models can explain the future behavior of macroscopic objects, such as glasses, rocks, and chairs, even if they do not include many of the variables that physically affect these systems. For example, when explaining the shattering of a glass of water, it is enough to show it as dependent on its falling from the table earlier. Many other variables such as the earth's mass, the behavior of air molecules, or absences of bursts of energy from outer space, likewise affect the glass's behavior. But these variables either make

only a tiny difference that does not show up in the coarse-grained variables we are interested in; or, they are almost constant, in which case we can take them into account without explicitly representing them. A model that says that the glass breaks if it falls from the table implicitly takes into account the effects of gravitation. However, we need not introduce an explicit variable for gravitation because it is almost always constant and so we do not need our model to still apply in circumstances where gravitational forces are drastically different.

Thus, rather than misrepresenting physical reality, our local causal models are explanatory because they represent robust dependencies between local variables. Moreover, robustness also has a straightforward connection to manipulation. If the dependence of *Y* on interventions on *X* is invariant, then it will hold in many different circumstances, and so agents like us can exploit the dependence even if we are ignorant of the exact circumstances. Our local models are thus successful because they represent objective physical dependencies between variables that have, due to their robustness, special relevance for explanation and manipulation.

5 Explanation and temporal directionality

Robustness in typical circumstance accounts for why many local models are explanatory.[70] At the same time, lack of robustness accounts for why some models are not explanatory.

Take a model that describes the later occurrence of the storm as dependent on the earlier reading of the barometer. This model is not explanatory because it is not invariant.

[70] In the following, when I talk about "robust," I always mean invariant and stable under the kind of circumstances that are typical or likely around here around now.

Suppose we intervene on the barometer reading by assuming an initial state just like the actual state, except that the barometer reading is different. If we evolve this new initial state forwards in accordance with the fundamental laws, we have no reason to think that it will change whether or not the storm occurs. The occurrence of the storm depends on the earlier atmospheric pressure in a large spatial region, and it is thus unlikely that a difference in just the reading of a barometer could create a force large enough to affect the storm. Therefore, the barometer reading is a mere record of the later storm, but it does not explain the storm because the dependence is not invariant under interventions.

I will now argue that the solution to the directionality challenge is that local backward-looking models are usually not robust. It thus makes sense for us to adopt time-asymmetric causal models because robust dependencies between local variables are widely available in the forward direction, but typically not in the backward direction. Although backward-looking models faithfully represent lawful dependence, lawful dependence in the backward direction is not equally explanatory because it is not robust between local variables.[71]

Why is there such an asymmetry? Many processes in our universe are irreversible in the sense that these types of processes never occur in time-reverse. For instance, glasses break, but glass pieces never spontaneously assemble into glasses; cigarettes burn to ashes, but ashes do not become cigarettes; rivers run down from mountains tops, but rivers never flow up onto mountain tops, etc. These processes are associated with an increase in entropy.

[71] We could find robust dependencies between highly gerry-mandered variables. But a general constraint on causal models is that the relevant variables need to be 'natural'. Moreover, models including gerry-mandered variables would be of no practical use.

The Second Law of Thermodynamics says that entropy tends to increase, but never decreases in the forward direction of time.

In irreversible processes, systems typically undergo a transition by which their energy gets more dispersed in space and where the system, as a consequence, becomes in some sense more disordered or less structured, such as breaking, being stirred, or cooling down. I will argue that models that depict earlier stages of such processes as dependent on later stages are either not invariant or not stable. So these models lack a central good-making feature of explanations, which rationalizes why they are not explanatory.

To not presuppose any time-asymmetry, I will rely on the time-neutral notion of a pure intervention introduced in section 3. Remember that in a pure intervention we take the entire state of the world at the relevant time and change only the intervened-on variable. All other variables at the time (inside or outside the model) are held fixed. We then use the laws of nature to evolve this state forwards and backwards to see the effect of this intervention on variables at other times.

Many backward-looking models are not invariant just as the barometer-storm model is not. The famous flagpole case is such an instance. The following equation relates the length of a flagpole to the length of its shadow, where *h* ranges over the length of the flagpole, *s* the length of the shadow a bit later, and *a* the angle of the elevation of the sun.

(F) h/s = tan(a)

However, while the earlier length of the flagpole at *t1* explains the later length of the shadow at *t2*, the length of the shadow at *t2* does not explain the length of the pole at *t1*. I

argue that we can account for this explanatory asymmetry because (F) is invariant under interventions on the length of the flagpole, but it is not invariant under interventions on the length of the shadow.

Start with the forward direction, where the length of the flagpole at *t1* explains the length of the shadow at *t2*. In intervening on the length of the flagpole at *t1*, we change only the length of the flagpole and hold everything else fixed. If we take this new state at *t1* and evolve it forwards in accordance with the fundamental physical laws, the length of the shadow is also increased at *t2* in accordance with (F). The reason is that some of the sunbeams that in the actual situation at *t1* pass the top of the flagpole and reach the area behind it, now, because of the added length of the flagpole, get blocked and do not reach this area. As a consequence, this area contains no photons at *t2*, which makes it shadowy (see FIG 2).

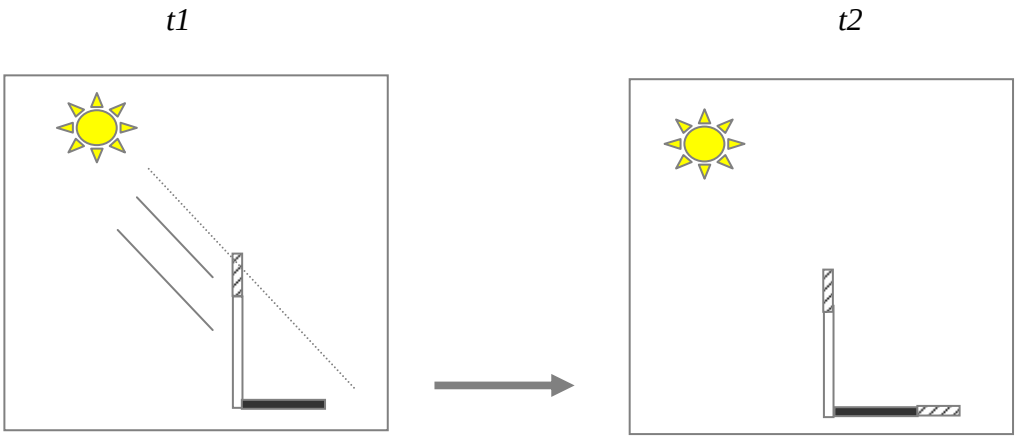

Figure 2 - *Flagpole Intervention*

We intervene on the length of the flagpole at *t1* by assuming a state where the flagpole is longer but everything else is the same. (The hatched area represents the added length of the flagpole at *t1*.) The shadow has its actual length at *t1* because we change only the length of the flagpole. However, because the intervention increases the flagpole's length at *t1*, some sun beams that actually pass the flagpole at *t1* and reach the area behind it at *t2* (represented by the dotted line) now get blocked by the added length of the pole. Contrary to what actually happens, these sunbeams thus do not reach a certain area behind the pole, making the shadow at *t2* longer than it actually is (represented by the hatched area). So the intervention increases the length of the shadow at *t2*.

In contrast, an intervention on the length of the shadow at *t2* does not change the length of the flagpole at *t1*. An intervention on the length of the shadow at *t2* increases the length of the shadow at *t2*, but leaves everything else at *t2* the same. The increase in the length of the shadow means that some region that actually has photons in it is now shadowy, and so contains no photons. Hence, the only difference between the actual state at *t2* and the intervened-on state concerns whether there are photons in some region behind the flagpole. In particular, the intervention does not change the flagpole's length at *t2*.

This intervention on the length of the shadow at *t2* does not change the length of the flagpole at *t1*. Here is the argument: The actual state of the world at *t2* (without the intervention) lawfully entails that the flagpole at *t1* has a certain length, say it is three meters long. The intervened-on state at *t2* differs from the actual state only with respect to the presence of some photons in a region behind the flagpole. But if the actual state at *t2* lawfully entails that the flagpole is three meters long at *t1*, then any state that matches the actual state at *t2* in all respects, except the presence of some photons, also entails that the flagpole is

three meters long. After all, the presence or absence of some photons could not generate the kind of energy or force necessary to shrink or expand a flagpole. Hence, an intervention on the length of the shadow at *t2* would not change the length of the flagpole at *t1* (see Diagram 2).

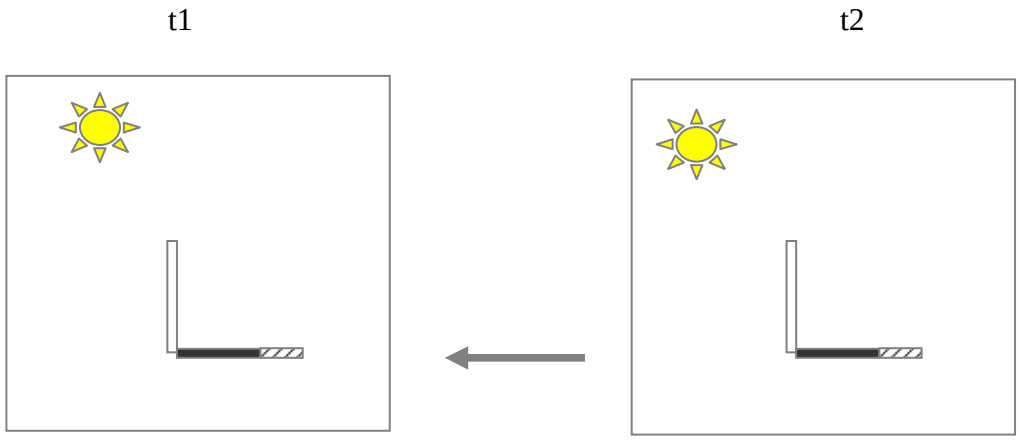

Figure 3 - *Shadow Intervention*

In intervening on the length of the shadow at *t2*, we assume a state of the world at *t2*, where the shadow is longer but everything else is the same. This change means that some photons have been subtracted from the area behind the flagpole, making the shadow longer (represented by the hatched area). Consequently, at *t1*, these photons are also missing. But there is absolutely no reason to think that the missing of these photons could generate a force large enough to change the length of the flagpole at *t1*.

In certain contexts we might assert backtracking counterfactuals of the form “If the length of the shadow were different, then the earlier length of the flagpole would have had to

have been different." But such counterfactuals do not show that (F) is invariant under interventions.

In evaluating the backtracking counterfactual, we reason about the most likely way in which a difference in the shadow's length could have been produced; and we might well think that such a scenario is one where the flagpole was shorter earlier. But such a scenario does not tell us about the outcome of an intervention at *t2* because in such a scenario other variables at *t2* are different as well. In particular, in a scenario where a longer flagpole at *t1* has produced a longer shadow at *t2*, this new state at *t2* differs from the actual state at *t2* not just with regard to the length of the shadow but also the length of the flagpole. Consequently, this scenario does not tell us about how earlier variables depend on changes in just the length of the shadow at *t2*.

So we can explain from the fundamental laws why there is an explanatory asymmetry in the flagpole case. The later length of the shadow does depend on interventions on the earlier length of the flagpole. A situation where the elevation of the sun and the current length of the shadow are the same, but the length of the flagpole is different would lawfully entail a difference in the length of the shadow a bit later. However, as shown, the length of the flagpole does not depend on interventions on the later length of the shadow. The relevant generalization is thus not explanatory because it is not invariant.

Many generalizations that allow us to infer the values of earlier variables from the values of later variables are like the flagpole case. For instance, in the case of photographs, memories, etc., the later variable is a mere record of some earlier variable. These generalizations are not invariant because interventions on the record-bearing state do not

generate a force large enough to change the recorded state, as the flagpole and barometer cases illustrate.

This explanation, however, does not work for all irreversible processes. Take a glass of water that is poised on the edge of a table, falls down, and shatters. Consider a model that represents the initial state of the glass on the table as dependent on the later location of the glass pieces on the floor. In contrast to the flagpole case, this model *is* invariant under interventions. An intervention that changes the later location of the glass pieces also changes the earlier state of the glass. A new final state, where the glass pieces are located differently, but everything else is equal, does not lawfully entail that a glass was on the table earlier.

This invariance is easiest to see if we visualize the process in time-reverse, i.e., as seen in a movie run backwards. We would see water flowing out from the cracks in the floor, while a glass assembles around it from numerous scattered pieces that also lie on the floor. Once assembled, the glass with the water in it jumps upwards from the floor and comes to rest exactly on the edge of the table (cf. Penrose 1989, 305).[72] This visualization shows that if the glass pieces were located differently on the floor, then even if each piece would still move the same way, the pieces would not assemble into a glass on a table earlier. Because the pieces would start out at different places, they would end up at different places and probably not form a glass at all.

However, despite this invariance, the model is not explanatory, because it lacks stability. As several philosophers have observed, in irreversible processes, the final state lawfully entails the initial macrostate only due to an extremely sensitive coordination among many components (cf. Dummett 1964, Elga 2000, Horwich 1987). Consequently, any small

[72] You can watch a glass-breaking in time-reverse in the following video: http://www.youtube.com/watch?v=NArwmQfUZ1M

interventions into the final state would disrupt this sensitive coordination and would make it that it no longer entails the initial macrostate (cf. Elga 2000). The dependence of the earlier state of a glass on the table on the later state where the glass pieces lie on the floor thus is not stable.

Again, this sensitivity is easiest to see if we imagine the process in time-reverse and think about what needs to happen for the glass pieces and water on the floor to assemble into a glass of water on the table. For the water to move out from the floor and for the glass pieces to start moving, billions of particles in the floor have to move in exactly the right way and bump into the pieces and the water molecules in just the right way to set them in motion. In addition, the movements of each glass piece and all the water molecules have to be coordinated with each other such that the glass pieces assemble around the water and the glass jumps onto the table. Moreover, the glass pieces have to be in the right molecular state to form chemical bonds upon collision.

So the initial macrostate of the glass standing on the table depends on the later state of the glass pieces lying on the floor. This dependence, however, is extremely unstable such that any small interference from the environment would disrupt it. The location of the glass pieces lawfully entails an earlier state of a glass on the table only due to its precise coordination with many other components, such as the movements of the particles in the floor, the movements of the water molecules, and the microstate of each glass piece.

If any component were slightly different, then the final state would no longer entail the actual macrostate even if the glass pieces have their actual location. Consequently, the dependence can also be disrupted by any external force that affects one of these components, such as small gravitational differences, colliding air molecules, or vibrations in the floor. So

there are no stable local models in such cases because any intervention on the background conditions makes it that the value of the earlier variable no longer depends on the value of the later variable.

This lack of stability makes the relevant models not explanatory. First, these generalizations do not isolate explanatory factors. The earlier state of the glass on the table depends on interventions on the later location of the glass pieces, but it equally depends on interventions on any other number of variables, such as the air pressure in the room, the state of the floor, and the exact microstate of each glass piece. It thus appears that the location of the glass pieces is not really responsible for the earlier macrostate, or at least no more so than numerous other variables.[73]

Second, it seems coincidental that so many independent variables have the very specific values required to lawfully entail the earlier macrostate of a glass standing on the edge of the table. Hence, by pointing to these variables, we do not remove the mystery of why there was a glass on the edge of the table earlier. In the forward direction, in contrast, we can isolate a small number of variables that lead to the later state of glass pieces on the floor irrespective of the exact values of many other variables. It might still be a coincidence that a fragile glass stood on a table just when someone pushed it. But at least the glass would still have shattered if it had been pushed a tiny bit stronger, if the table had been a little bit higher, or if there had been a bit of wind.

[73] Hausman (1998) gives a similar account of the explanatory asymmetry in these cases. He focuses on the idea that earlier variables on which an outcome depends are typically statistically independent, whereas later variables on which it depends are statistically correlated. This proposal is similar to my robustness proposal in that it entails that the relevance of a later variable to an earlier outcome is extremely difficult to determine. The reason in his case is that a later variable cannot be changed without also changing all the other later variables with which it is correlated. See Hausman's book for an insightful discussion of this proposal and how it relates to other theories.

Third, because the values of the variables have to be so specific in the backward-looking case, we get no interesting unification. If we look at two different cases in which we find glasses on a table and then glass pieces on the floor, we need different models to account for each case because the explanation is sensitive to the exact details of the case. These details, however, will vary among different cases. Fourth, the model does not help us for manipulation or prediction. Because the dependence is so sensitive, we cannot determine whether a system is in that state, nor could we intentionally prepare it to be in that state.

So in cases of irreversible processes there are no robust generalizations in the backward direction because either stability or robustness fails. Hence, there are no local, backward-looking models that can support explanations. Moreover, the majority of macroscopic processes in our universe are irreversible.

But what about processes that are not irreversible? Suppose a billiard ball moves from one end of a billiard table to the other end. This process is reversible because an identical ball could traverse the same path in the opposite direction. In such cases, we can build backward-looking models that are both invariant and stable. Take the billiard ball's momentum at *t1* and its momentum at *t2*. There are invariant generalizations about how the earlier momentum of the ball at *t1* depends on its momentum at *t2*. Moreover, these generalizations are robust because they hold for a wide range of background conditions, such as different temperatures or precise locations of the ball. So why can we not explain the earlier state of the ball in terms of its later state? Moreover, the same question arises for other reversible processes.

I think there are several replies here. One reply is that the model is explanatory though the explanation is not very good or interesting. Demands for causal explanation

usually arise when systems undergo significant changes or deviate from the state we would expect them to be in. For example consider this passage by Hart and Honoré:

> The notion that a cause is essentially something which interferes with or intervenes in the course of events which normally take place, is central to the commonsense concept of cause ... Analogies with the interference by human beings with the natural course of events in part control, even in cases where there is literally no human intervention, what is identified as the cause of some occurrence; the cause, though not a literal intervention, is a difference to the normal course which accounts for the difference in outcome. (Hart and Honoré 1985, 29)

Reversible processes are typically such that a system does not undergo significant macroscopic changes. If Hart and Honoré are right, then a process where a billiard ball moves over a table would not need to be explained causally because the system does not deviate from its normal course. After all, a billiard ball that keeps moving with constant velocity is just what we would expect. So one reason why we do not adopt these backward-looking causal models, even when invariant generalizations are available, is that we do not feel that the relevant phenomena are in need of explanation.

I want to put more weight on a second and third reply though. The second reply is that explanations are cumulative. Part of our explanatory practice is that a given model can typically be situated within a more extensive model. For instance, in our ordinary way of explaining things, I might have a model that says that a ball moves with a certain momentum because it already had that momentum a second earlier. Now, I can go further back in time and say that the ball had this momentum a second earlier because someone poked it with a

stick three seconds earlier. If I explain things in terms of earlier things, I am typically able to situate explanations within more extensive explanations.

But we could not do that if we explain things in terms of later things. Say the billiard ball falls into a pocket in the future. You can then not give a more extensive explanation of why the ball had the momentum it had earlier in terms of future occurrences because falling into a pocket is an irreversible process; the ball's lying in the pocket does not explain its earlier momentum due to lack of robustness. Explanations that we could give in terms backward-looking models would be extremely isolated in that we could not relate them to more extensive explanations. The general scarcity of robust generalizations in the backward direction is thus a reason not to adopt backward-looking models even in the few cases where robust generalizations are available.

The third reply is that, as I will show in the next section, backward-looking models do not guide us in interacting with the physical world, not even in cases where dependencies are robust. An important part of causal explanations is that they are recipes for how to manipulate the world. But explanations of earlier variables in terms of later variables could not fulfill that important purpose. The fact that backward-looking causal models are typically not robust accounts for why it makes sense for us to not adopt these models for the purpose of explanation.

6 Agency and temporal directionality

Another main reason why we are interested in forward-looking causal models is their action-guidance. As we have seen, in many cases interventions on later variables lawfully entail changes in earlier variables just as interventions on earlier variables lawfully entail

changes in later variables. Yet, we can often use earlier variables as means for changing later variables, but never *vice versa*. Hence, forward-looking, but not backward-looking, models can guide our interactions with the world. I will argue that we can explain why there is this asymmetry from the structure of fundamental physics.

Suppose an isolated billiard ball moves across a frictionless table. We can build a forward-looking model where the later momentum of the ball at *t2* depends on interventions on its earlier momentum at *t1*. For instance, if an intervention changes its momentum at *t1* to *p*, then its momentum at *t2* also changes to *p*. This model is action-guiding because it tells me that preparing an otherwise isolated billiard ball to have momentum *p* at *t1* is a means for preparing it to also have momentum *p* at *t2*. For instance, if I give the ball a push shortly before *t1* to change its momentum at *t1* to *p*, I thereby also change its momentum at *t2* to *p*. (Taking into account friction and other intervening factors will make a difference, but the model is still approximately true.)

Now, we can also build a backward-looking model where the earlier momentum of the ball at *t1* depends on interventions on its later momentum at *t2*. But if I give the ball a push shortly before *t2* to change its momentum at *t2* to *p*, I thereby do not also change its earlier momentum at *t1* to *p*. So the backward-looking model does not guide me in my interactions with the world in the same way as the forward-looking model does. This curious feature needs explanation given that the later momentum of the ball lawfully determines its earlier momentum just as much as *vice versa*.[74] In a recent paper, Mathias Frisch argues that

[74] The important point here is not so much that the actual fundamental laws are deterministic in both temporal directions (though there is good reason to believe they are), but that this asymmetry of agency arises even *if* the laws are deterministic in both temporal directions.

the best explanation of this asymmetry of agency is the existence of primitive, time-asymmetric causal facts over and above lawful entailment (cf. Frisch 2010).

Pace Frisch, I argue that we can explain this asymmetry (why forward-looking, but not backward-looking, models are action-guiding) from physical facts about our make-up as agents. The crucial point is that changes brought about by human actions are different from the "pure interventions" that I have been considering so far. In a pure intervention into the final state of a system at a time, we assume some variable changed but hold everything else fixed. So we compare the actual final state of the system to a copy where some variable is changed but everything else is equal. Our causal models (both backward- and forward-looking ones) tell us about how variables change given such pure interventions into either the initial or the final state of a system.

Agent-interventions (i.e., changes to a system brought about by an agent), however, are different from pure interventions in that we also have to take account of the agent. We have to compare the actual final state to a copy that differs not only with respect to the changed variable but also with respect to the state of the agent. For instance, in a scenario where I change the momentum of the billiard ball at *t2* by performing an intentional action (for instance, by pushing it), not just the state of the ball is different from its actual state at *t2*, but the state of my body is also different from a situation in which I have not acted (for instance, some momentum will have left my body).

Importantly, agent-interventions are time-asymmetric because we exercise agency via intentional actions and intentional actions are time-asymmetric. In paradigmatic cases, intentional actions involve deliberation or some other decision-making procedure. So if I change some variable by virtue of an intentional action, there must be some process

associated with my action that involves deliberation and decision-making. Agents like us deliberate and decide *before* they act. For instance, if I intentionally change the momentum of the billiard ball at *t2*, then I have deliberated and decided before *t2*. So an asymmetry about agents like us is that in intentional action we intervene into a system from the past.

A consequence of this psychological asymmetry is that systems into which we agent-intervene behave toward the future approximately as they would behave given a pure intervention, but they do not behave toward the past approximately as they would behave given a pure intervention. My forward-looking model of the billiard ball tells me that an intervention that changes its momentum at *t1* to *p*, if the system remains isolated, also changes its momentum at *t2* to *p*. In intentional action, I first deliberate and then push the billiard ball shortly before *t1*. Because my action does not interfere with the evolution of the ball between *t1* and *t2*, the ball behaves just as it would under a pure intervention.

Now consider an agent-intervention on its momentum at *t2*. My backward-looking model tells me that a pure intervention that changes the ball's momentum at *t2* to *p* also changes its momentum at *t1* to *p*, given that the ball remains isolated between *t1* and *t2*. However, if I change the ball's momentum at *t2* by virtue of an intentional action, then the system is not isolated between *t1* and *t2*. My very own action interferes with the ball's isolated evolution toward the past. In a situation where I have pushed the ball shortly before *t2*, the ball is not isolated between *t1* and *t2*. You can see this by picturing the process of my pushing the ball in time-reverse. In this scenario, the momentum of the ball is not preserved towards the past as it would be if the system were isolated (in which case the ball's momentum would also be *p* at *t1*), but some of the momentum of the ball, instead of being preserved in the ball, travels from the ball to my hand. So in an agent-intervention the system

interacts with its environment in a way in which it does not interact with its environment in a pure intervention. Because agent-interventions have this interfering effect on the past evolution of a system, knowing how the history of a system would change given a pure intervention cannot guide our actions.

The outstanding puzzle is then why there is this psychological time-asymmetry. Primitivists about causation might still argue that we need a primitive causal asymmetry to explain why agents like us deliberate in this time-asymmetric way. Why are we set up such that we deliberate *before* our actions? Presumably, the explanation is that there is evolutionary pressure for agents to deliberate in this time-asymmetric way (Pearl 2000, 59). But which features of the physical structure of our world create this pressure?

A constraint on deliberation and intentional action is that the agent must be in a position to assume that her decision will lead to the respective action. Agents who deliberate about whether to decide to φ or to not-φ, but whose decision would not be correlated with them φ-ing or, respectively, not φ-ing, would be very ineffective. Effective agents must be in a position to assume that their decisions are correlated with the body movements that they decide to perform. In fact we can be very certain that our decisions to perform basic actions result in the corresponding body movements.

But this certainty requires that there are robust generalizations connecting our decisions to the requisite actions. Our decisions have to be correlated with the respective body movement for a wide range of decisions within a wide range of circumstances. Otherwise, we would not be in a position to assume that the appropriate body movement will happen since circumstances vary widely across decision situations. Now, we have seen from the discussion of explanation that such robust connections are rare in the backward direction

but common in the forward direction. Robust connections are more likely to be found between an agent's decisions and future body movements rather than past body movements. Thus, it makes sense that effective agents have evolved to deliberate for future outcomes but not past outcomes, and so only forward-looking models guide our actions.[75]

7 Conclusion

In this paper, I give an account of why our ordinary causal assumptions make sense in light of the world presented to us by fundamental physics. I argue that our local, time-asymmetric causal models represent robust dependencies between variables. And the scarcity of robust dependencies between local variables in the backward direction accounts for why only forward-looking models support explanation and action-guidance.

[75] I give a more detailed account of the time-asymmetry of control in chapter 2 of this dissertation.

REFERENCES

Cartwright, N. (1979) "Causal Laws and Effective Strategies," *Nous* **13**: 419-437.

Dowe, P. (2000) *Physical Causation*. Cambridge: Cambridge University Press.

Dummett, M. (1964) "Bringing about the Past," *Philosophical Review* **73**: 338-359.

Eagle, A. (2007) "Pragmatist Causation," in: Price and Corry (2007), 156-190.

Earman, J. (1976) "Causation: a matter of life and death," *Journal of Philosophy* **73**: 5-25.

Elga, A. (2000) "Statistical Mechanics and the Asymmetry of Counterfactual Dependence," *Philosophy of Science* **68** (Supplement): 313-324.

Elga, A. (2007) "Isolation and Folk Physics," in: Price and Corry (2007), 106-119.

Field, H. (2003) "Causation in a Physical World," in: Loux, M. and D. Zimmerman (eds.) *Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 435-460.

Friedman, M. (1974) "Explanation and Scientific Understanding," *Journal of Philosophy* **71**: 5-19.

Frisch, M. (2005) *Inconsistency, Asymmetry, and Non-Locality*. Oxford: Oxford University Press.

Frisch, M. (2010) "Causal Models and the Asymmetry of State Preparation," in: Suárez, M., M. Dorato, and M. Réde (eds.) *EPSA Philosophical Issues in the Sciences, vol.2*. Springer Press, 75-86.

Hart, H. and Honoré, T. (1985). *Causation in the Law*, 2nd edn. Oxford: Oxford University Press.

Hall, N. (2004) "Rescued From the Rubbish Bin: Lewis on Causation," *Philosophy of Science* **71**: 1107-1114.

Hall, N. (2005) "Causation," in: Jackson, F. and M. Smith (eds.) *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press, 505-533.

Hausman, D. (1998) *Causal Asymmetries*. Cambridge: Cambridge University Press.

Healey, R. (1983) "Temporal and Causal Asymmetry," in: R. Swinburne (ed.) *Space, Time, and Causality*. Dordrecht: D. Reidel Publishing, 79-103.

Hitchcock, C. (2007) “Prevention, Preemption, and the Principle of Sufficient Reason,” *Philosophical Review* **116**: 495-532.

Hitchcock, C. (2009) “Causal Modeling,” in: Beebee, H., C. Hitchcock, and P. Menzies (eds.) *The Oxford Handbook of Causation.* Oxford: Oxford University Press, 299-314.

Horwich, P. (1987) *Asymmetries in Time*. Cambridge: MIT Press.

Kitcher, P. (1989) “Explanatory Unification and the Causal Structure of the World,” in: Kitcher, P. and W. Salmon (eds.) *Scientific Explanation*. Minneapolis: University of Minnesota Press, 410-505.

Lange, M. (2000) *Natural Laws in Scientific Practice.* Oxford: Oxford University Press.

Lange, M. (2009) *Laws and Lawmakers. Science, Metaphysics and the Laws of Nature.* Oxford: Oxford University Press.

Lewis, D. (1986a) “Counterfactual Dependence and Time’s Arrow,” in: *Philosophical Papers*: *Volume II*. Oxford: Oxford University Press, 32-52.

Lewis, D. (1986b) “Causation,” in: *Philosophical Papers: Volume II*. Oxford: Oxford University Press, 159-213.

Lockwood, M. (2005) *The Labyrinth of Time*. Oxford: Oxford University Press.

Loewer, B. (2007) “Counterfactuals and the Second Law,” in: Price and Corry (2007), 293-326.

Loewer, B. (2012) “Two Accounts of Laws and Time,” *Philosophical Studies* **160**: 115-137.

Maudlin, T. (2007) *The Metaphysics Within Physics*. Oxford: Oxford University Press.

North, J. (2008) “Two Views on Time Reversal,” *Philosophy of Science* **75**: 201-223.

Norton, J. (2007) “Causation as a Folk Science,” in: Price and Corry (2007), 11-44.

Owens, D. (1992) *Causes and Coincidences*. Cambridge: Cambridge University Press.

Papineau, D. (1985) “Causal Asymmetry,” *British Journal for the Philosophy of Science* **36**: 273–289.

Penrose, R. (1989) *The Emperor's New Mind*. Oxford: Oxford University Press.

Price, H. (1996) *Time’s Arrow and Archimedes’ Point*. Oxford: Oxford University Press.

Price, H. (2007) “Causal Perspectivalism,” in: Price and Corry (2007), 250-292.

Price, H. and R. Corry (2007) *Causation, Physics and the Constitution of Reality*. Oxford: Oxford University Press.

Reichenbach, H. (1956) *The Direction of Time*. Berkley: University of California Press.

Reutlinger, A., Schurz, G. and Hüttemann, A. (2011) "*Ceteris Paribus* Laws," *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/ceteris-paribus/>.

Russell, B. (1913) "On the Notion of Cause," *Proceedings of the Aristotelian Society* **13**: 1-26.

Schaffer, J. (2010) "Review of Price and Corry's *Causation, Physics, and the Constitution of Reality," Mind* **119**: 844-848.

Tooley, M. (1987) *Causation: A Realist Approach*. Oxford: Oxford University Press.

van Fraassen, B. (1993) "Armstrong, Cartwright, and Earman on Laws and Symmetry," *Philosophy and Phenomenological Research* **53**: 431-444.

Weslake, B. (2006) "Review of *Making Things Happen*," *Australasian Journal of Philosophy* **84**: 136-140.

Woodward, J. (2003) *Making Things Happen*. Oxford: Oxford University Press.

Woodward, J., and C. Hitchcock (2003) ''Explanatory Generalizations, Part I: A Counterfactual Account,'' *Nous* **37**: 1–24.

Woodward, J. (2007) "Causation with a Human Face," in: Price and Corry (2007), 66-105.