CAUSATION IN A PHYSICAL WORLD

BY THOMAS BLANCHARD

A dissertation submitted to the
Graduate School—New Brunswick
Rutgers, The State University of New Jersey
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
Graduate Program in Philosophy

Written under the direction of
Barry Loewer
and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2014

ABSTRACT OF THE DISSERTATION

Causation in a Physical World

by Thomas Blanchard

Dissertation Director: Barry Loewer

This dissertation offers a new solution to the problem of causation in the physical world. Fundamental physics leaves little space for causation. Causation is local and asymmetric, but physical laws are global and time-symmetric. However, causal notions are indispensable. In particular we need causation to make sense of effective strategies. The problem of causation in the physical world is the challenge of reconciling the a-causal physical picture of the world with the need for causation. Chapter 1 describes the problem in detail and proposes a new methodology to solve it. The proper method to handle the problem isn't conceptual analysis. Rather, solutions to the problem should be judged on how well they physically explain actual facts about effective strategies. Chapter 2 examines the main attempts to solve the problem. I argue that they all face various problems. In particular, current attempts to locate causation in the physical world all have trouble making sense of the fact that we need causal knowledge to make rational decisions. To solve this problem, the first step is to provide a satisfactory explanation of why only those correlations that are (intuitively) causal can be exploited for the purpose of securing desired outcomes. In chapter 3, I propose such an explanation. I argue that causal correlations are the only ones that can be exploited according to *evidential decision theory* (EDT). This is a surprising claim, since EDT is widely thought to recommend acting for the sake of outcomes one cannot cause. I argue

that this is actually not the case, and that EDT in fact provides a plausible account of exploitable correlations. In chapter 4, I use this account to offer a new solution to the problem of causation in a physical world. I argue that causal dependence is a matter of the cause and the effect standing in certain probabilistic relations to a third event called a *probabilistic intervention*. Probabilistic interventions are events that need not involve agency but nonetheless mimic certain crucial features of deliberation. I argue that this account provides a plausible solution to the problem of causation in a physical world.

Acknowledgements

First and foremost, I would like to thank my adviser Barry Loewer and my committee members Branden Fitelson, Jenann Ismael and Jonathan Schaffer.

I cannot overstate how much I owe to Barry and how much influence he has had on my philosophical development. I took the first of many seminars with Barry during my first semester of grad school. Although the content of the seminar went largely over my head, it is where I became deeply interested in the metaphysics of science, especially the tensions between the scientific and the manifest image and the problem of asymmetries in time. By the end of that semester, I knew I wanted to work under Barry's supervision. As should become obvious upon the most cursory examination of this dissertation, his way of thinking about issues in the philosophy of science pervades my entire research. In many respects, this dissertation is a footnote to Barry's work. Barry's constant guidance, support and friendliness have been invaluable on both an academic and a personal level. I am extremely grateful to have had him as adviser.

My specific interest in causation traces back to a wonderful seminar on Woodward's book *Making Things Happen* taught by Branden Fitelson at Rutgers in the fall of 2010. Branden has taught me a great lot about interventionism, decision theory and formal methods of doing philosophy. In addition, I am very grateful for his unfailing support and enthusiasm over the last few years.

I first met Jenann Ismael when she visited Rutgers a couple of years ago. Since then, I have benefited enormously from talking with her and reading her deeply original work about Humeanism, causation and decision. I am also grateful for her constant friendliness and interest in my project. I am fortunate to have had an external committee

member as generous with her time, attention and support as Jenann.

I owe a huge debt to Jonathan Schaffer. It goes without saying that I have learned an enormous lot about causation from Jonathan. But he has also taught me much about philosophical method in general and how to become a professional philosopher. Over the last years I haven't ceased to be amazed by his incredible generosity with his time and attention. (I cannot remember how many of my drafts he has suffered through.) His challenging questions and comments have greatly helped me clarify my project and ideas. His practical help and moral support during my job search has been invaluable. In addition, Jonathan kindly offered to have me as a co-author for a paper on actual causation.[1] This has been an extremely useful learning experience for me, and I am very grateful for it.

I am also very grateful to David Albert for his constant interest in my project. Our numerous conversations about causation, time-asymmetry and chance, as well as his challenging questions, have greatly helped me shape my thoughts. I have also received very helpful feedback from Laurie A. Paul, Huw Price and Michael Strevens. I have learned a lot about causation from them, and I thank them for taking the time to read various portions of my material and discuss it with me.

I owe special thanks to my friends Heather Demarest and Michael Hicks. Over the last six years, I have benefited immensely from our regular conversations on various topics in the philosophy of science, from their friendship and from their constant support. Harjit Bhogal, Alison Fernandes, Christian Loew and Zee Perry also provided very helpful comments on parts of the material presented here. I would also like to thank audiences at the University of Delaware, Illinois Wesleyan University, Rice University, the 2011 British Society for the Philosophy of Science Annual Conference and the Prague First Ernst Mach Colloquium.

I would also like to thank the other people with whom I have discussed the ideas presented here or have otherwise helped me during the dissertation process: Zack Al-Witri, Nick Beckstead, Mary Coleman, Mark Criley, Marco Dees, Raphael Ehrsam, Andrew

[1] Blanchard T. and Schaffer, J. (2014). Cause without Default. In Beebee, H., Hitchcock, C., and Price, H. (Eds.) *Making a Difference*. Oxford: Oxford University Press, forthcoming.

Engen, Melinda Fagan, Alvin Goldman, Erik Hoversten, Douglas Husak, Emily Kelahan, Philippe Lusson, David Rose, Kurt Sylvan, Christopher Weaver, Tobias Wilsch and Dean Zimmerman. I hope I have not unwittingly failed to mention anyone.

I gratefully acknowledge the financial support of the Rutgers University Graduate School (New Brunswick), the philosophy department at Rutgers, the John Sellon Charitable Trust and the Mellon Foundation.

Last but not least, an enormous thanks to my parents and sister for their support over the (sometimes challenging) last six years, and to my wife Audrey. I wouldn't have been able to finish this dissertation without her love and unlimited support.

Contents

Chapter 1

The Problem of Causation in a Physical World

Causation is a central ingredient of reality as we understand it. A concern for causes pervades our lives: we wonder why the air conditioning is not working, why our friend seemed so gloomy today, whether our career choices will lead to happy lives. Moreover, the central goal of the special sciences (biology, medicine, economics, history, and so on) is to discover the causes of phenomena of interest. Practitioners of those disciplines are concerned with questions such as why dinosaurs became extinct, what are the factors responsible for heart attacks, whether deflation causes unemployment, what caused the Civil War, etc. Finally, the notion of causation plays a central role in our best accounts of many important philosophical notions, such as reference, knowledge, perception, and so on. As John Carroll says, 'with regard to our total conceptual apparatus, causation is at the center of the center' (1994, 118). Thus, understanding the nature of causation is a central task for philosophy of science, metaphysics and philosophical anthropology.

This dissertation attempts to make progress on this task by focusing on what Hartry Field has called 'the central problem in the metaphysics of causation' (2003, 443). This is the *problem of causation in a physical world* or *Russell's problem*, named in honor of the first philosopher who explicitly discussed the underlying issues. In a nutshell, the problem is as follows. In his famous article 'On the Notion of Cause' (1913), Russell

pointed out that physics seems to leave no space for causation: our best physical theories of the world, he argued, are incompatible with the existence of causal relations. Russell thus advocated *causal eliminativism*, the view that there are no causal relations in our world. But as we will see, positing the existence of causal facts is indispensable to accomplish certain crucial explanatory tasks. This leaves us with the challenge of reconciling what fundamental physics tells us about the nature of our world with the need to posit causal relations. My intent in this dissertation is to offer a systematic and comprehensive solution to this puzzle, one that fares better than other solutions currently on the market.

In this chapter, my goal is to explain in more detail what Russell's problem is, and what solving it involves exactly. In §1, I summarize Russell's argument for the claim that fundamental physics leaves no space for causal relations. I then discuss two reactions one may have to Russell's argument. The first one is simply to accept Russell's conclusion and endorse causal eliminativism (§2). The other is to deny Russell's assumption that causal facts (if there are any) are grounded in fundamental physics (§3). Both options, I will argue, are implausible. The only remaining option is to try to articulate a theory of causation that explains how the physical structure of our world can (contrary to what Russell claimed) give rise to causal facts. I close the chapter by clarifying the goals of such a theory and the methodology we should follow in trying to develop one (§4). There I will argue that Russell's problem calls for a strong methodological reorientation in the metaphysics of causation. Specifically, I will argue that the proper methodology to solve Russell's problem is not the traditional method of conceptual analysis. Rather, the task of a proper solution to Russell's problem is to provide a satisfactory physical explanation of actual facts about effective strategies. I will also argue that the kind of causal relation (often called 'actual causation') on which philosophers of causation have tended to concentrate is not the one that matters in the context of Russell's argument. Rather, the relation that matters is *difference-making* or *causal dependence*. These methodological preliminaries will pave the way for the critical examination of current solutions to Russell's problem in chapter 2.

1 Russell's (1913) Attack on Causation

In 'On the Notion of Cause' (1913), Russell launched a forceful attack on causation. As he puts it in the famous introduction to his piece,

> All philosophers, of every school, imagine that causation is one of the fundamental axioms or postulates of science, yet, oddly enough, in advanced sciences such as gravitational astronomy, the word 'cause' never occurs. Dr. James Ward... makes this a ground of complaint against physics... To me it seems that... the reason why physics has ceased to look for causes is that, in fact, there are no such things. The law of causality, I believe, 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. (1913, 1)

Russell's attack has many targets. Hitchcock (2007) finds four distinct theses in this passage:

T1. The notion of cause is incoherent.
T2. There are no causes in our world.
T3. The 'law of causality' is false.[1]
T4. The word 'cause' should be expunged from philosophical vocabulary.

Hitchcock further notes that although Russell doesn't clearly distinguish between these theses, they are not equivalent. **T1** does entail **T2**, **T3** and arguably **T4**, but **T3** doesn't entail **T2** or **T1**. Even if not every event has a cause, this doesn't mean that there are no such things or that our concept of cause is fundamentally confused. Likewise, the non-existence of causes doesn't entail the incoherence of the notion of cause. Finally, **T4** neither clearly entail nor is clearly entailed by either **T2** or **T3**. Hereinafter I will concentrate on Russell's defense of causal eliminativism (**T2**). His case for **T2** rests in part on his argument for **T1**, which is that the main definitions of causation on offer at the time of his writing have insuperable problems. This argument is unconvincing, since as Hitchcock (2007, 50) points out the absence of an adequate definition of a concept doesn't entail the concept's incoherence. But Russell also offers another, much more powerful argument for eliminativism that doesn't presuppose that our notion of cause is

[1] The law of causality is the postulate that every actual event has a cause that necessitates its occurrence. As Mill puts it: 'The law of causation, the recognition of which is the main pillar of inductive science, is but the familiar truth that invariability of succession is found by observation to obtain between every fact in nature and some other fact which has preceded it' (1916, Bk. III, ch. 5, §2). Russell offers a convincing argument against the law of causality: see Hitchcock (2007, 47-8) for a reconstruction of this argument, and Norton (2007a) for a contemporary attack on the law of causality and related causal postulates that is very much in Russell's spirit.

confused. The gist of Russell's argument is that (a) the existence of causes presupposes that the fundamental physical laws of our world have certain features but (b) our best fundamental physical theory of the world - for Russell, classical mechanics - reveals that the laws lack the relevant features. Our belief in causation is a 'relic of a bygone age' insofar as it relies on an outdated and misguided conception of the fundamental physical laws of our world. How exactly the argument is supposed to go is not entirely transparent in Russell's text, and to reconstruct it in its best form we will have to make use of some contemporary philosophical tools and distinctions unknown at the time of Russell's writing.[2] Along the way I will argue that the strength of Russell's argument is left largely unaffected by developments in physics that occurred after 1913. Even if classical mechanics has now been superseded by better competitors for the status of fundamental physical theory of the world (namely relativity theory and quantum mechanics), Russell's argument still has bite.

Russell's argument presupposes that causation is at bottom a matter of physics. This view, which I will call *physicalism about causation*, can be formulated more precisely as follows:

> **Physicalism about Causation**. If there are causal facts, they either are or reduce to actual fundamental physical facts.

On this view, either causation appears in the fundamental physical ontology of our world, or it is non-fundamental and can somehow be reduced to the elements of this ontology.[3] The sense of 'reduction' at play here is *ontological* reduction: **Physicalism about Causation** says nothing about the reducibility of causal *concepts* to more basic concepts. Rather, it says that if causal facts are not fundamental they are somehow *grounded in* or *dependent on* or *metaphysically explained* by fundamental physical facts.[4] Note that physicalism about causation entails that causal facts supervene on

[2]My reconstruction of Russell's argument largely follows Field's (2003) influential interpretation of Russell. See also Eagle (2007, 156-162), who points out certain crucial unstated premises of Russell's argument.

[3]These two possibilities are not exclusive. It could be that some causal facts are fundamental physical facts while other, higher-level causal facts reduce to the fundamental physical level.

[4]There are deep issues about how exactly to explicate the concepts of grounding, metaphysical dependence and metaphysical explanation, but I won't go into these issues here. I will simply trust that the reader has some ideas of what these notions mean.

fundamental physical facts: fix the fundamental ontology of our world, and the causal facts are thereby fixed. Russell's argument for causal eliminativism has two main parts. First, he argues that the fundamental physical ontology of our world doesn't contain causal relations (§1.1). Second, he argues that there is nothing in the actual fundamental physical structure of our world to which causation can be reduced (§2). Given **Physicalism about causation**, the upshot of these two arguments is that there are no causes in our world, because causes cannot be located anywhere in physics.

1.1 Causation Is Not Part of Fundamental Physics

To determine whether there are causal relations in the fundamental physical ontology of our world, the natural strategy is to take our best candidates for the status of fundamental physical theories and see if they posit primitive causal relations. If we abstract away from the details of these theories, we can see that they posit three sorts of facts. First, there are facts about the geometric structure of spacetime. Second, there are facts about the material content of spacetime: that is, facts about which fundamental physical properties or fields (mass, charge, and so on) are instantiated at each spacetime point, and/or facts about the existence of point-like particles. Third, there are facts about the physical laws that govern or describe the distribution of material contents within the spacetime arena. Note that in the metaphysics of physics there is an important debate about whether facts about laws are really fundamental physical facts. According to Humeans (e.g. Lewis (1983), Loewer (1996)), facts about laws of nature reduce to facts about material contents of the universe: laws of nature are simple, informative statements about the distribution of fundamental properties across spacetime points. Anti-Humeans (e.g. Armstrong (1983) Dretske (1977), Tooley (1977), Maudlin (2007)) hold that laws are physically fundamental: laws are entities over and above the material contents of our universe which govern the distribution of physical stuff in the spacetime arena. In what follows I will remain neutral on this debate, as it doesn't bear on the issues I will discuss.[5]

[5]One exception: as we will see the brand of anti-Humeanism defended by Maudlin is relevant to some of the issues raised by Russell's argument that the causal time-asymmetry cannot be grounded in fundamental physics.

Russell assumes that if causation is a primitive physical relation, it will appear in the physical laws - i.e. that the best statements of those laws will make use of the word 'cause'. But it is worth examining first whether the two other sorts of facts posited by our best fundamental physical theories may be primitive causal facts. It seems clear that facts about the geometric structure of spacetime are not primitive causal facts. Almost none of our best fundamental physical theories make use of causal vocabulary to describe geometric relations between spacetime points. The only exception is the causal set approach to quantum gravity developed by Bombelli et al. (1987) and Reid (2001).[6] On this approach, time is represented as a growing block whose development is underlain by primitive *causal* relations between points. If this theory is correct then *contra* Russell the fundamental physical ontology comprises primitive causal facts. But causal set theory is far less developed and popular than the two leading approaches to quantum gravity (string theory and quantum loop theory). So the existence of causal set theory isn't a very powerful argument against Russell's contention that causation isn't a fundamental physical relation. Turning to facts about the material content of the universe: it also seems relatively clear that those sorts of facts are not primitively causal. There are two main metaphysical theories of fundamental physical properties, categoricalism and dispositionalism. If the former is true, the fact that (e.g.) a point-particle has a certain mass is an intrinsic fact, not a relational one and *a fortiori* not a causal one. According to dispositionalism, it *is* a relational fact, since on this view the property of having a certain mass is individuated by the nomological relations between mass and other physical properties. But then the question whether facts about fundamental properties are primitive causal facts boils down to the question whether physical laws describe primitive causal relations. Russell convincingly answers the latter question in the negative. As he points out, causation doesn't appear as a primitive in the laws of classical mechanics. Consider for instance Newton's second law, $F = ma$. This is simply a differential equation relating rates of change in various quantities at a time; it doesn't say that those quantities are causally related.[7] Moreover,

[6] See Earman (2008, sect. 7) for a good summary of causal set theory.

[7] Note that Humeans and anti-Humeans disagree on how best to formulate the second law of classical mechanics (and other candidates for the status of fundamental physical law). Humeans think that

this is plausibly true for the laws of other fundamental physical theories as well. It is sometimes argued that special and general relativity rely on primitive causal laws. For instance, in general relativity one finds a principle of 'local causality' (Hawking and Ellis, 1973, 60). But as Norton (2007b) points out when spelled out this principle just amounts to the requirement that the (classical) fields at a spacetime point be entirely fixed by the fields in its past light-cone. Thus one doesn't really need the word 'cause' to formulate the principle. Norton convincingly argues that the same holds true of other causal principles that appear in certain formulations of relativity theory and quantum mechanics. All in all, then, Russell's claim that causation isn't a primitive physical relation appears very convincing.

1.2 Causation is Not Reducible to Fundamental Physics

The fact that causation doesn't appear in fundamental physics isn't enough to establish that causation doesn't exist. Causal facts may well be non-fundamental facts whose existence can be explained in terms of (i.e. reduced to) the non-causal material one finds in fundamental physics. In the second part of his argument, Russell argues that in fact, the actual fundamental physical structure of our world cannot ground the existence of a (non-fundamental) relation of causation. The upshot is that not only is causation absent fundamental physics, it is not even implicitly contained in the picture of reality given to us by fundamental physics.

Russell's argument goes as follows. First, he assumes that causation requires *physical determination*, where c physically determines e just in case c together with the physical laws entails e.[8] Call this view **Determination**:

it is best expressed as a statement describing the following regularity: whenever the forces exerted on an actual corpuscle are such-and-such, the mass and acceleration of the corpuscle are such and such. (Clearly, no primitive causal relation appears in this statement.) By contrast, anti-Humeans will insist that the second law is best regimented as a statement describing necessitation relations between the universals *force*, *mass* and *acceleration* (Armstrong, 1983; Dretske, 1977; Tooley, 1977) or as a statement describing a primitive relation of governance between mass and forces on the one hand and acceleration on the other hand (Maudlin, 2007). One may have the suspicion that 'necessitation' and 'governance' are other names for a primitive relation of causation. But as we will see in §1.2, necessitation and governance (if they exist) do not deserve to be called 'causation' because they lack some of the central features of causation; in particular, in our world they cannot hold between localized (spatially small) events.

[8] Note that entailment is a relation between propositions. For convenience, throughout this dissertation I use lower-case italic letters to designate both an event and the proposition that the event occurs. No

Determination. c causes e just in case c physically determines e.

Moreover, Russell assumes that causation has the following two features:

Localization. Causes sometimes are localized (i.e. spatially small) events.[9]
Asymmetry. Causation is asymmetric.

(I will explain these two features in more detail shortly.) Moreover, Russell assumes that **Determination**, **Localization** and **Asymmetry** are all there is to say about the nature of causation. That is, causation *just is* asymmetric, localized physical determination. Correspondingly, to believe that there is causation in our world just is to believe that the physical determination relations that drive the evolution of our world are asymmetric and localized. Russell then goes on to argue that if our best physical theories of the world are correct, there are no such physical determination relations in our world: actual physical determination relations hold between global states of the world only, and they have no asymmetry built into them. Our belief that causation exists is a 'relic of a bygone age' insofar as it relies on an obsolete and misguided picture of the physical laws of our world.

I'll explain Russell's arguments for the claim that actual physical determination relations are neither localized nor asymmetric in §**1.2.1** and §**1.2.2** respectively. Beforehand let me make three remarks on **Determination**. First, the idea that causation reduces to physical determination isn't peculiar to Russell. It has a historical antecedent in Hume's famous theory of causation, which makes it a necessary condition for c causing e that c and e be of constantly conjoined types. Since for Hume laws of nature are just regularities (so that it is a law that es follow cs just in case every instance of c is followed by an instance of e) this amounts to the claim that c causes e only if c together with some law of nature entails e. Add to this the physicalist claim that all events are physical events (so that laws relating events are *ipso facto* physical) and you get **Determination**.

harm should come from this ambiguity. In general, 'c' is intended to represent a putative cause, and 'e' a putative effect. I will use capital italic letters to designate event-types or variables.

[9] I use the term 'localization' rather than 'locality' so as to prevent any confusion with the notion of locality as it is used in relativity theory.

Second, although as we will see **Determination** should be rejected, the view of causation it encapsulates is not without pre-theoretical plausibility. There is a tight connection between causation and laws of nature that any plausible theory of causation must recognize and accommodate. In particular, causal and nomological relations can both be pre-theoretically described as relations of *necessitation* and *production* that drive the development of our world through time. The fact that both relations are naturally describable with the same vocabulary points to a deep link between them. By making nomological determination necessary for causation, **Determination** provides a very simple and precise articulation of this connection.

Second, on the conception of causation embodied in **Determination**, the existence of causal facts requires determinism. For an event c together with the physical laws to entail the occurrence of a distinct event e, the relevant laws must be deterministic. Russell took no issue with this, since like everybody else at the time of his writing he took classical mechanics to be a deterministic theory.[10] But since the advent of quantum mechanics we know that the fundamental laws guiding the physical evolution of our world may be irreducibly stochastic. On the standard, collapse interpretation of quantum mechanics, the state of the world at this time doesn't determine whether (say) a particular kaon atom will decay in the next five minutes. The laws only fix a non-trivial probability (i.e. one strictly between 0 and 1) for its decay. If this standard interpretation (or some other collapse interpretation like GRW) is correct, this may appear to give us a much quicker route to causal eliminativism than the one followed by Russell, for if the laws are stochastic then there are no physical determination relations in our world and thus given **Determination** no causal facts either. There is, however, a natural response to this line of thought. To accommodate the possibility of stochastic laws, one may replace **Determination** by the weaker requirement that a cause c together with the laws of nature must entail a certain *objective probability* of occurrence for its effect e. (Determinism then becomes a special case in which the objective chance of the effect occurring is 1.) Call this weakening of **Determination**

[10] We now know that this is false: see Earman (1986) and Norton (2008). In what follows I will leave aside the complications raised by Earman's and Norton's discussions.

Probabilistic-Determination. This view preserves the close connection between causation and the laws encapsulated in **Determination** while accommodating indeterminism. As we will see, Russell's considerations about localization and asymmetry provide an argument against the existence of causation even if one replaces **Determination** by **Determination-Probabilistic**.

1.2.1 Localization

Localization, remember, is the idea that localized events can be and sometimes are causes, where a localized event is (as Russell puts it) 'something short of the whole state of the universe' (1913, 7) at a time. The centrality of this feature of causation is attested by the fact that paradigmatic examples of causes are spatially small events. If we had to produce a paradigmatic case of a causal relation, we would presumably cite the throwing of a rock causing the window to break, the scratching of a match causing the forest to burn, the cue ball causing the 8 ball to sink, and so on.

Russell points out, however, that localized events are not of the right kind to enter into the physical determination relation: a localized event by itself is never sufficient to physically determine what we intuitively regard as its effects. For instance, Suzy's throwing of the rock by itself doesn't nomologically determine the window to shatter. Obviously, it is nomologically possible that Suzy's throw fail to be followed by the window breaking, for instance if the rock is deviated from its trajectory by a strong gust of wind, or is intercepted by Billy along the way. Whether or not the throw at t_0 is followed by the window-breaking at t_1 nomologically depends on many other factors besides what happens at t_0 in the region of the throw. And the example generalizes to any other instance of causal relation involving localized events: there is always the possibility of an outside interference that prevents the cause from bringing about its effect. This means that no putative physical law relating localized events has any chance of being true. And indeed in classical mechanics, determining what happens at some time t_1 requires information about the *complete* state of the world at t_0. Since classical mechanics puts no constraints on the speed at which influence may travel, to nomologically determine what happens in some spatial region at t_1 one must specify what happens

in all regions of spacetime at t_0. Even assuming, in line with special relativity, that influence cannot travel faster than light, anything less than a specification of everything that happens inside the entire cross-section of the window breaking's past light-cone at t_0 will fail to determine the window-breaking (or an objective chance for it) at t_1. If the two times are more than a few nanoseconds apart the relevant cross-section will be a spatially enormous event.

On Russell's point of view, this has two consequences. First, it means that physical determination relations are not properly characterized as causal, since they don't take localized events as inputs. Moreover, given **Determination**, it also entails that localized events cannot be causes. Since **Localization** is a central feature of causal relations, this means that there is nothing in nature that satisfies our concept of cause.

1.2.2 Asymmetry

The second central feature of causation targeted by Russell's argument is the asymmetry of causation. Although Russell doesn't make the distinction explicitly there are two different phenomena that fall under this heading and that it will be important to keep separate in what follows. The first is the fact that causation has a direction or an 'arrow':

> **Direction**. If c causes e, then e is not (or at least not generally) a cause of c.

The parenthetical hedger is required because there might be cases of symmetric causation. One standard example involves two planks of wood standing against each other to form an upside-down V. One may well think that each plank standing in a certain position is a cause of the other standing in a certain position.[11] Whatever one thinks of this case, however, it is clear that in a great majority of actual cases the causal relation runs in one direction only.

The second, related phenomenon is the fact that causation is time-asymmetric:

[11] Menzies (1989) replies that in this example it is really the position of one plank *at a certain time* that causes the position of the other at *a later time*, so that there is in fact no symmetry once one individuates the causal relata properly.

Time-Asymmetry. Causes normally precede their effects.

That is, the causal arrow typically points in the same direction as the temporal arrow. The directionality of causation, note, doesn't entail its time-asymmetry. One can imagine a world in which every cause has effects in both directions of time but effects are never causes of their causes. But the time-asymmetry of causation presupposes that causation has a direction. If effects routinely caused their causes, there would be no predominant temporal direction of causation. **Direction**, note, says only that causes *normally* precede their effects. The need for this hedger arises from the fact that backward causation may be possible in our world under exceptional circumstances, as the equations of general relativity allow for the possibility of closed time-like curves. For all we know there might be such curves in our world, in which case it would be possible for (say) the behavior of a particle at a time t to affect the particle's position at an earlier time.

Some physicists and philosophers have also argued for a retro-causal interpretation of EPR correlations (see e.g. de Beauregard (1977)). It is an open question whether this interpretation of EPR correlations is true. If it is, then there is backward causation in our world. Nevertheless, there is no doubt in familiar circumstances involving macroscopic objects, causal processes go from past to future.[12]

Russell argues that physical determination relations are neither directed nor (*a fortiori*) time-asymmetric. As he puts it, physics shows us that 'the supposed lack of symmetry between 'cause' and 'effect' is illusory' (1913, 11). His case rests on the fact that the laws of classical mechanics are *bi-directionally deterministic*. This means that in classical mechanics the complete state of the world at a certain time not only physically determines the complete state of the world at any later time; it also determines the complete state of the world at any earlier time. Bi-directional determinism entails that there are plenty of pairs of states such that each physically determines the other. Indeed, for any two pairs of actual complete states of the world at some time

[12]Note that 'normally' in the statement of **Time-Asymmetry** doesn't mean *infrequent*. If the retro-causal interpretation of EPR correlations is correct, there is plenty of backward causation in our world. Rather, 'normally' means something like *in circumstances with which we as laypeople are familiar*.

S_t and $S_{t'}$ (with t' later than t), not only does S_t determines $S_{t'}$ but the reverse also holds. This means that physical determination doesn't satisfy **Determination**: there are plenty of cases where the determination relation between two states goes both ways. It also means that physical determination is not time-asymmetric, since later states of the world determine earlier states.[13] So nomological relations are not of the right kind to be counted as causal relations. Moreover, on the assumption that **Localization**, **Asymmetry** and **Determination** collectively exhaust the content of our concept of cause, the only possible source of the causal direction and time-asymmetry is a direction and time-asymmetry in the physical laws themselves. Since there is no such thing, this means once again that nothing in the world that can ground asymmetric causation.[14]

Here as in other parts of Russell's argument, there is the question whether these considerations still have bite when we move to physical theories that have superseded classical mechanics. Deterministic interpretations of quantum mechanics (Bohmian mechanics and the many-worlds theory) are also bi-directionally deterministic, so that Russell's argument still applies to them. Issues become more complicated when we move to indeterministic theories. As Field (2003) points out, Russell's argument need not rely on determinism. If the true physical theory of the world is indeterministic, Russell could still run the same argument (using **Probabilistic-Determination** instead of **Determination**) as long as on this theory the state of the world at some time determines a probability distribution not only over later states but also on earlier states. But as Field also notes, in actual indeterministic interpretations of quantum mechanics the physical laws have an asymmetry built into them. For instance, in the GRW interpretation of quantum mechanics, the laws fix a probability distribution over

[13] As Russell says: '[T]he future 'determines' the past in exactly the same sense in which the past 'determines' the future' (1913, 15).

[14] It is sometimes said that Russell's argument appeals to the fact that classical mechanics is *time-symmetric* - i.e. applying a time-reversal operator to a sequence of states physically allowed by classical mechanics yields a sequence of states that is also allowed by the theory. Whether or not a theory is time-symmetric is independent of whether it is bi-directionally deterministic. Farr and Reutlinger (2013) convincingly argue that time-symmetry cannot do the job done by bi-directional determinism in Russell's argument. This doesn't mean it isn't relevant to the debate about the sources of the causal asymmetry. We will see in the next chapter that the time-symmetry of classical mechanics (and other physical theories) can be leveled to raise troubles for an influential account of the causal asymmetry due to Lewis.

the world's evolution toward the future but do not say anything about its evolution toward the past. Albert (2000, ch. 7) argues that if GRW is correct thermodynamic and related physical asymmetries might be explained in terms of this nomological asymmetry; perhaps it could help explain the time-asymmetry of causation as well. We shouldn't put much weight on this possibility, however, for reasons that will appear in section 4.2.2 of this chapter. A satisfactory explanation of the asymmetry of causation, I will argue there, should be compatible with all or most of our main candidates for the status of a fundamental physical theory. This means it should be compatible with fundamental theories that do not comprise any nomological asymmetry.

1.3 Summary

To summarize, Russell's argument for causal eliminativism is as follows. If causation exists, it is either part of the fundamental physical structure of our world, or it somehow reduces to it. But causation isn't part of the fundamental physical ontology; fundamental physics doesn't contain primitive causal relations. So if causation exists, it must be grounded in a-causal fundamental physical facts. For the purpose of his argument Russell endorses the following physicalist reductive view of causation: causal relations are just localized, asymmetric physical determination relations. But fundamental physics shows us that there are no such relations in our world. So physics leaves no space for the existence of causal relations in our world. Our causal beliefs are simply false beliefs about the character of the physical laws.

What should we make of this argument? Russell's case for the claims that causation isn't fundamental and that there are no localized, asymmetric physical determination relations in our world is very plausible, and I will assume the correctness of these claims in what follows. This leaves us with three main options. The first one is simply to accept Russell's argument, and endorse causal eliminativism. The second one is to maintain causal realism but reject the first premise of Russell's argument - the claim that causal facts are either fundamental physical facts or reduce to them. That is, one may insist that causation is real but is an extra-ingredient of reality over and above physics. I will call this view *anti-physicalism about causation*. The last option

is to accept physicalism about causation and the claim that causation isn't part of fundamental physics, but nonetheless maintain the existence of localized, asymmetric causal relations in our world. Clearly such a view must deny **Determination**, the idea that causation requires physical determination.[15] In the next two sections, I will review the first two options and argue that they are unattractive. This leaves us with the third option. The main challenge there is to articulate a plausible physicalist alternative to the view of causation proposed by Russell - one that doesn't lead to eliminativism. Such an alternative theory would show how the existence of localized asymmetric causal relations can be grounded in the non-causal structures and relations one finds in fundamental physics. In the last section of this chapter, I will discuss what such a theory should accomplish exactly, and the methodology we should adopt in trying to develop it.

2 Causal Eliminativism

The first stance one might adopt towards Russell's argument is simply to accept its conclusion that causation has no place in a proper scientific understanding of the world.[16] But there is a decisive objection against causal eliminativism due to Cartwright (1979). Her argument against causal eliminativism is an indispensability argument for the existence of causation: we need to posit the existence of localized, (time-)asymmetric causal relations to explain certain obvious facts about *effective strategies*. Later on we

[15] Two remarks here. First, note that to maintain that there is real *asymmetric* causation in our world, one need not in principle reject **Determination**. Rather, what one needs to reject is Russell's assumption that **Determination**, **Localization** and **Asymmetry** exhaust the content of our concept of cause. This assumption is essential to Russell's asymmetry argument since from it it follows that the only possible source of the causal asymmetry is an asymmetry in physical determination. One could reject the remark by maintaining that causation is physical determination *plus* some asymmetric relation between cause and effect. Such a view would not solve the problem of localization, however.

The second remark is that one could also in principle maintain physicalism and causal realism by endorsing the view that **Localization** and **Asymmetry** are not essential features of causation. On this view what Russell's argument would actually show is not that there are no causes, but that in our world causation is global and symmetric. Ney (2009) comes close to endorsing such a view. We'll see in the next section that this view doesn't solve the real challenge raised by Russell's argument, however.

[16] Note that although Russell endorsed this view in 1913, later on in his life he came to abandon it. In his (1948) he offers a realist theory of causation (an early version of what is now called the physical process theory of causation).

will spend quite some time unpacking what the idea of an effective strategy amounts to exactly. For now, we may say that the phrase 'effective strategies' stands for the pre-theoretical idea that one may reliably achieve a desired goal by performing a certain action. It is obvious that in our world certain actions are more effective strategies than others to accomplish a desired goal. For instance, scratching a match is a good way to create a fire, while dunking the match in water isn't. Cartwright's point is that we need to posit the existence of causation to explain why certain actions are good strategies and others are not: as she puts it, the existence of causation 'grounds the distinction between effective strategies and ineffective ones' (1979, 420). For instance, intuitively scratching a match is an effective strategy for creating a fire *because* scratching a match is a cause of fire. By contrast, dunking the match in water isn't effective *because* it is not causally conducive to the occurrence of a fire. Another example: manipulating the reading on a barometer isn't an effective strategy to influence the occurrence of a storm later, despite the fact that a low reading on the barometer is often followed by a storm. The obvious reason is that the barometer reading doesn't cause the occurrence of the storm; rather, both are effects of a common cause (low atmospheric pressure).

Note that the two features of causation targeted by Russell's argument - **Localization** and Asymmetry - are essential to explain our abilities to achieve certain goals. Were there no localized, asymmetric relation of causation in our world, we couldn't explain certain striking facts about effective strategies for goal advancement. On the one hand, **Localization** is essential to explain why we can sometimes achieve a desired result by performing a certain action. The kinds of actions we can perform - striking a match, throwing a rock, and so on - are localized events. If localized events could not be causes, actions could not be effective strategies for anything. On the other hand, **Asymmetry** - that is, **Direction** and **Time-Asymmetry** - is needed to explain certain striking and uncontroversial general features of effective strategies. The first is that generally causes are effective strategies for influencing the occurrence of their effects, but effects are not effective strategies for influencing the occurrence of their causes. (To avoid lung cancer, it is a good idea to stop smoking, but if the goal is to stop smoking it is useless to take a cancer-preventing pill (Field, 2003).) This

presupposes that causation has a direction - i.e. that effects are not in general causes of their causes. The second is that effective strategies have a striking *temporal orientation toward the future*. Among the present actions at my disposal there are many through which I can usefully influence the future, but none by which I can usefully influence the past. The natural explanation of this fact is that in our world causation is always or almost always future-directed. Thus, insofar as it leads us to reject the existence of localized and asymmetric relations of causation in our world, causal eliminativism is unacceptable since it deprives us of a crucial explanatory resource to make sense of the existence and general characteristics of actual effective strategies.[17] Without causation, we cannot distinguish effective from ineffective strategies, and we cannot explain why we can influence the future but not the past.

Cartwright's argument is certainly a powerful reason to reject causal eliminativism. But why think it is the strongest one? Another argument against causal eliminativism is that although the concept of causation isn't needed in fundamental physics, it is an essential tool for other corners of the scientific inquiry, as witnessed by the fact that appeal to causation is ubiquitous in the special sciences. Special scientific theories are couched in causal terms, and researchers in the special sciences are chiefly concerned with discovering causes (e.g. of the Civil war, of heart attacks, of the extinction of dinosaurs, and so on). The argument, then, is that causal eliminativism is intolerably costly because it makes our best special scientific theories of the world false and implausibly entails that the main objective of researchers in those sciences is fundamentally misguided. However, this argument doesn't strike me as very powerful (or rather: insofar as it has force, it is precisely because of the indispensability of causation for explaining effective strategies). Consider the claim that causal eliminativism makes our best special scientific theories false. The causal eliminativist might respond that even if we stripe our best scientific theories of their causal content, they still give us

[17] By the same token, so does the view mentioned in the preceding footnote, on which Russell's argument shows that causation is actually a global, symmetric relation. If this is all that causation is in our world then it cannot help us explain the localized, directed and future-directed character of effective strategies; insisting that we should nonetheless call it 'causation' is beside the point. Indeed, Ney (2009), who comes close to endorsing such a view, recognizes the need to posit causal relations that satisfy **Localization** and **Asymmetry** to explain effective strategies.

important information about our world. Even if we deny the existence of causation, we can still recognize that our world is full of *correlations*. For instance, we can deny that increased demand *causes* increased supply but still recognize that there is a correlation between increased demand and increased supply. That is: episodes of increased demand for a good are followed in higher proportion by higher supply for this good than episodes of decreased demand. The causal eliminativist might say that even if the causal generalizations of economics are false, economic theory is still valuable insofar as it discovers interesting, non-obvious correlations of this sort. So causal eliminativism doesn't force us to implausibly contend that the special sciences have nothing of value to teach us about the world. (The causal eliminativist cannot say this about effective strategies, since a mere correlation between an act and an outcome isn't sufficient for the act being an effective strategy to achieve the outcome. The symmetric fact that (e.g.) smoking is correlated with lung cancer cannot explain why quitting smoking is an effective strategy for avoiding lung cancer but taking cancer-preventing pills is an ineffective strategy to quit smoking.) Now, it is true that researchers in the special sciences do not simply care about correlations. For instance, econometricians are deeply concerned with the question whether increasing the money supply has positive effects on unemployment and economic growth. But this is due in large part to the fact that econometricians are concerned with determining whether increasing the money supply is an effective policy for increasing economic growth. Econometricians need causal concepts because to distinguish effective from ineffective strategies we need causal and not only correlational information. One may suspect that this is true for other special sciences as well.

One may reply that researchers in the special sciences are concerned with causation (and not only correlation) because the project of those sciences is not only to discover effective strategies but to *explain* phenomena. For instance, medical biologists are concerned with explaining heart attacks, and to do so they need information about the *causes* of heart attacks. More generally, one might think that causal eliminativism should be rejected because it cannot make sense of our explanatory practices. We think that the height of the flagpole explains the length of its shadow but not the

reverse, and the intuitive reason is that the former is a *cause* rather than an *effect* of the latter. According to this argument, causal eliminativism is intolerably costly because it cannot make sense of our (asymmetric) explanatory practices. But the point of our explanatory practices isn't entirely clear in the first place. That is, it isn't entirely obvious why we care about explanation, and (consequently) why abandoning our ordinary explanatory practices would be intolerably costly. The only theory I know of that explains the importance of explanatory information is due to Woodward (2003). On this view, explanatory information is important because it is information potentially relevant for manipulation and control. The length of the flagpole explains its shadow (but not the reverse) because manipulating the flagpole's height is an effective strategy for manipulating the shadow (and not the reverse). This makes it clear why we care about explanation: explanatory information is information about effective strategies for attaining goals, and we have an obvious interest in acquiring knowledge about what would be best conducive to the realization of our goals. But if this view is correct, the argument that causal eliminativism is intolerably costly because it cannot make sense of our explanatory practices derives its force from Cartwright's argument. It seems to me, then, that Cartwright's appeal to effective strategies is clearly the strongest argument against causal eliminativism.[18]

3 Anti-Physicalism about Causation

Another position one might adopt in light of the tension between fundamental physics and causation pointed out by Russell is to endorse anti-physicalism about causation, the view that causal facts are neither fundamental physical facts nor grounded in the latter. On this view, causation is simply an extra-ingredient of reality over and above the realm of facts described by fundamental physics. This position solves the tension between the indispensability of causation and Russell's arguments by maintaining that

[18] Yet another argument against causal eliminativism is that the notion of causation is essential to make sense of other philosophically important concepts such as knowledge, reference, disposition and so on. However, the causal eliminativist might well maintain that causation isn't indispensable for understanding these concepts: perhaps the notion of correlation is enough. For instance, it is not implausible to think that for one to know that p, it is not necessary that one's belief that p be caused by p. Perhaps it is enough that the belief be reliably correlated with p.

causes are real but denying that the causal aspect of the world can be reduced to its physical aspect. This is the view endorsed by Cartwright (1979) in reaction to Russell's arguments.[19] Anti-physicalism about causation is also entailed by causal anti-reductionism, the more general view that causal facts do not ontologically reduce to non-causal facts (physical or otherwise). Causal anti-reductionists often motivate their view by appealing to Russell. For instance, Carroll (2009) mentions Russell's argument against the possibility of grounding the causal asymmetries in symmetric physics as a motivation for anti-reductionism.[20]

There are two main problems for this view. The first one is that causal anti-reductionism threatens to make causal facts epistemologically inaccessible. The concern here is that if causal facts do not supervene on physical facts, causal claims are *under-determined* by all physical evidence that is in principle available. But plausibly we can only observe physical sequences, which makes it hard to see how we could come to know the truth-value of causal claims if anti-physicalism about causation were true. There are two replies one could make to this argument (Schaffer, 2009). First, one might object that underdetermination arguments in general lead to unacceptable skepticism. If the aforementioned argument were correct, then by the same token one could argue that one cannot have knowledge of the external world since its existence is underdetermined by the totality of our experience. But the two situations are not analogous, I think. In the latter case, a natural answer is that the existence of the external world *best explains* the particular features of our experience, so that we are warranted to believe in it by inference from the best explanation. It is not clear at all that causal facts can do similar explanatory work if anti-physicalism is true. After all, physical sequences can presumably be entirely explained in terms of the material one finds in fundamental physics, so that non-physical causal facts are explanatorily superfluous. A second reply the anti-physicalist might make is that contrary to what the underdetermination argument presupposes we have direct experience causal facts, as argued by e.g. Fales (1990, ch. 1) and Armstrong (1997, 211-16). However, the scope of Fales's and Armstrong's

[19] Field (2003) calls this view *hyperrealism*.
[20] For a detailed defense of causal anti-reductionism, see Tooley (1987, 1990).

arguments is quite limited: they argue that we have direct experience of causal facts when we perceive pressure on the body and when we experience the operations of our will. In other cases of causation the claim that we have direct experience of the relevant causal facts is far less plausible. So the argument does nothing to explain how we can come to know that (e.g.) deflation causes unemployment or that smoking causes cancer.

A second problem for anti-physicalism about causation is that it also threatens to make causal facts practically irrelevant. Since anti-physicalism about causation denies the supervenience of the causal on the physical, it seems committed to there being a possible world that is exactly identical to our world in all physical respects, but in which (for instance) smoking doesn't cause cancer, so that in this world refraining to smoke isn't an effective strategy to avoid lung cancer. Even though for the reasons pointed out by Russell it is hard to see how facts about effective strategies can be grounded in physics, it seems even more implausible to think that facts about effective strategies can float entirely free from the physical nature of our world.

4 Solving Russell's Problem: Methodological Issues and Stage-Setting

Let me summarize where we are so far. As we have seen, Russell's arguments give us strong reasons for thinking that fundamental physics leaves no space for causation in our world. But causal eliminativism has catastrophic consequences, as it deprives us of a crucial resource for making sense of effective strategies. And positing a causal realm over and above physics does little to solve the problem. The only remaining option is to maintain that causes exist (*pace* Russell) but that their existence is grounded in fundamental physics (against Cartwright and others). At a minimum, this requires rejecting the crucial premise of Russell's argument for eliminativism, the assumption that causation is a form of physical determination. But simply rejecting **Determination** isn't enough to alleviate the worries raised by Russell's argument. As I noted when I introduced **Determination**, any plausible view of causation that rejects **Determination** must nevertheless maintain a close connection between causation and physical

laws. And given the globality and symmetry of physical laws, it remains mysterious whether and how the laws leave room for causal relations in our world. I will call this predicament the *problem of causation in a physical world* or *Russell's problem*. A satisfactory solution to the problem should show how the fundamentally a-causal physical structure of our world can nevertheless ground the existence of localized, asymmetric causal relations.

In chapter 2 I will critically examine existing attempts to solve Russell's problem, and present a new one in chapters 3 and 4. But first, I will discuss certain metaphilosophical and methodological issues about what a proper solution to Russell's problem should achieve exactly. Although many solutions to Russell's problem have been proposed, there has been little discussion in the literature of what a solution to the problem should look like and of the desiderata it should satisfy.[21] Getting clear on those issues will be very helpful when we turn to the task of reviewing existing solutions to the problem and devising a new one. It will also go at least some way toward alleviating the following worry. Solving Russell's problem may seem to require providing a *conceptual analysis* of causation. But the conceptual analysis of causation appears to be a hopeless endeavor, as witnessed by the fact that despite enormous philosophical efforts in the last decades no successful analysis of causation has yet been provided. I will argue that solving Russell's problem doesn't require giving a conceptual analysis of causation, and that the reasons for being skeptical of conceptual analysis do not apply to the project of solving Russell's problem.

'Conceptual analysis of causation' is an umbrella phrase that applies to a wide variety of projects. Nevertheless, one can distinguish two central tenets endorsed by all or most of its practitioners. First, conceptual analysts of causation all aim to complete the schema

(S) c causes e iff...

in non-causal terms and in a way such that the resulting sentence is true in all possible

[21] Kutach (2013, ch. 1) is an exception. The methodology I will argue is best appropriate for solving Russell's problem is close to the one Kutach proposes. But see fn. 22 below.

worlds, not just the actual one. In other words, a successful conceptual analysis is supposed to have a very wide modal scope. Second, conceptual analysts of causation are committed to a particular procedure of evaluation. A proposed completion of **(S)** should be evaluated chiefly in terms of how well it fits our *intuitions* about what causes what in various hypothetical cases. These two commitments are rather broad and can be cashed out in different ways. Regarding modal scope, some philosophers (e.g. Ducasse (1926)) aim for a *definition* of the concept of causation: a conceptually necessary completion of **(S)** whose right hand side includes only concepts that are (a) more basic than the concept of causation and (b) must be possessed by anyone capable of using the word 'cause' competently. Contemporary conceptual analysts of causation (for instance Paul and Hall (2013)) tend to be less concerned with our concept of causation than with causation in the world, and aim for a *metaphysically* necessary analysis. Regarding the question of criteria of evaluation, there is room for various views about how tight the fit with intuitions should be and about the relative importance of other criteria of evaluation. Contemporary philosophers of causation tend to agree that an analysis need not fit *all* our intuitions to be successful, that it may involve revision of our ordinary concept of causation, and that it should also be evaluated in terms of standard scientific criteria for theory evaluation (e.g. simplicity).

The projects of solving Russell's problem and of conceptually analyzing causation are similar in certain respects. A solution to Russell's problem should show how causal relations are grounded in fundamental physics: ideally this should take the form of a completion of **(S)** whose right-hand side refers only to elements of the fundamental physical ontology. Since the latter doesn't include causal relations, this means that like a conceptual analysis a proper solution to Russell's problem should be a reduction of causation. But the reduction need not have the status of a necessary truth (§4.1). Moreover, a proposed solution to Russell's solution should be judged not on how well it accommodates causal intuitions, although intuitions have a role to play. Rather, it should be judged on how well it explains effective strategies - or so I will argue in §4.2.[22]

[22]The methodology I propose is similar to the one developed by Kutach in his book *Causation and its Basis in Fundamental Physics* (2013, ch. 1). Kutach also argues that the proper task of a metaphysics of causation is to articulate a physically grounded concept of causation that can help us

4.1 Causation in the Actual World

Consider the question of the modal scope of **(S)** first. The most pressing issue raised by Russell's arguments is that fundamental physics seems to leave no room for the existence of causation *in our world*. Correspondingly, a solution to Russell's problem need only explain what causation is *actually*, in a way that makes its existence compatible with what physics tells us our world is like. In that respect, the project isn't conceptual analysis but what Dowe calls an 'empirical analysis' of causation, whose goal is to 'establish what causation in fact *is* in the actual world' (2000, 3). As we will see empirical analysis may require looking at other possible worlds, as consideration of close-by alternatives is relevant to the question of the relations between causation and physics in our world. But a successful solution to Russell's problem need not apply to very distant worlds such as worlds with outlandish laws of nature. This is an important respect of difference with conceptual analysis. Conceptual analyses of causation often fail because they give the wrong verdicts in worlds very different from ours, e.g. worlds with laws of magic. Thus Tooley (1987), Carroll (1994) and Schaffer (2000b, 2001) provide a battery of hypothetical examples involving magical laws against various proposed conceptual analyses. Given the nature of our project we can simply ignore these examples when evaluating proposed solutions to Russell's challenge. It is no strike against a theory of causation in the actual world that it fails to capture alleged causal facts in possible worlds very distant from ours.[23]

explain the phenomena which account for the utility of our folk concept of causation, in particular the phenomenon of effective strategies. There are important differences between our two approaches, however. For one thing, Kutach says little about how to determine what counts as an effective strategy in the first place. My own account gives an explicit role to intuitions about practical rationality in determining what is an effective strategy for what. Second, Kutach says little about what explaining effective strategies amounts to. I will discuss the explanatory virtues that a proper physical account of effective strategies should satisfy below.

[23] Here is one potential worry with this argument. Earlier I said that a solution to Russell's problem should show us how fundamental physics can ground the existence of causation, and grounding facts are widely taken to be metaphysically necessary. Doesn't it follow, then, that a completion of **(S)** that solves Russell's problem must have the status of a necessary truth? Not so. If a sentence of the form 'c causes e iff X' exhibits the physical facts that ground causation, it follows that in any world where X obtains the corresponding causal fact also obtains. But this is compatible with the claim that in worlds very distant from ours (such as magical worlds) there are relations that deserve the name 'causation' but do not reduce to X.

4.2 Explaining Effective Strategies

Let's turn now to the question of the criteria one should employ to evaluate proposed solutions to Russell's problem. Instead of using fit with causal intuitions in hypothetical cases as the methodological cornerstone, I propose that attempts to solve to Russell's problem be evaluated in terms of *how well they explain actual facts about effective strategies*. As we will see this criterion of evaluation leaves some role for intuitions, but there is more to explaining than merely accommodating those intuitions.

This criterion of evaluation is appropriate because as we saw the best reason to posit causal facts in our world is precisely that those facts are necessary to explain effective strategies. For instance, the fact that throwing rocks causes windows to break explains why if you want the window to break it is a good idea to throw a rock at it. Causal facts also explain *general* facts about effective strategies. The fact that localized events (including human actions) can be causes explain why we can sometimes act so as to achieve desired outcomes. Likewise, the fact that causation is future-oriented explains why we can sometimes usefully influence the future but we can never influence the past. This means that to a satisfactory solution to Russell's problem should show that, contrary to what Russell claims, there is a physical relation that can explain these facts about effective strategies. That is, suppose that a proposed solution to Russell's problem identifies causation with some physical relation R. Then for the solution to be satisfactory, it should be the case that (e.g.) the fact that rock-throwing and window-breaking stand in relation R explains why the former is an effective strategy for the latter; that the fact that the barometer dial and the storm do not stand in relation R explains why manipulating the barometer dial isn't an effective strategy to make a storm occur; that the fact that R is future-directed explains why we can influence the future but not the past, and so on. (So at a minimum R should be able to have localized events such as rock-throwings as relata, and it should be time-asymmetric and hence have a direction. That is, it should have those features which Russell argued cannot be found in fundamental physics.)

To make this criterion of evaluation sufficiently precise, I will say a few words about

what *explaining* effective strategies amounts to in this context (§§4.2.1-4.2.3). It will also be useful to unpack the notion of effective strategy a bit further (§4.2.4).

4.2.1 Extensional Adequacy

To frame the task at hand as an *explanatory* one is to say that proposed solutions to Russell's problem should be judged on how much they display the various virtues that explanations can have. An important explanatory virtue (although not the only one) is what I will call *extensional adequacy*. Suppose we identify causation with some physical relation R. If R is to explain (e.g.) why throwing rocks is a good strategy to break windows, it better be the case that R holds between cases of rock-throwing and cases of window-breaking. More generally, if R is to explain effective strategies, it should be the case that whenever an action is an effective strategy for a desired outcome, R holds between the action and the outcome. This is simply applying to the case at hand a general constraint on good explanations: they should be able to capture the range of phenomena that they are designed to explain. For instance, a good statistical-mechanical explanation of Boyle's law should be such that when the relevant statistical-mechanical facts hold, the relations between temperature, pressure and volume encoded in Boyle's law should also hold.

Extensional adequacy means intuitions do have a role to play in theory evaluation. If a proposed solution to Russell's problem entails that an action c is a cause of an event e but intuitively c is not an effective strategy for e, this should count against the theory. More generally, intuitions can help us delimit the range of phenomena that a proper solution to Russell's problem should explain. But note that the relevant intuitions are intuitions about what is an effective strategy for what in our world, not intuitions about causation *per se*. As we will see this is an advantage since our intuitions about effective strategies display a remarkable degree of clarity and intersubjective agreement.[24] Moreover fit with intuitions doesn't play the overarching role it has in conceptual analysis, for extensional adequacy isn't the only virtue that explanations can display.

[24] See §4.2.1 below.

4.2.2 Other Explanatory Virtues

Let me point out three other explanatory virtues that will play an important role in the following chapters.

A first important explanatory virtue is *generality* or insensitivity. An explanation is better the more insensitive to details it is. Insensitivity can be cashed out as counterfactual robustness: an explanation is insensitive to certain respects in which the world is just in case if the world were different in these respects, the facts which account for the *explanandum* would still hold. To illustrate consider Putnam's (1975) famous example of the square peg and the round hole. We can explain why a solid rigid square peg with a 1-inch diagonal doesn't fit into a round hole with a diameter of 1 inch in terms of rigidity, solidity and geometry, or by mentioning the specific elementary particle constitutions of a particular metal square peg and wooden board. The former explanation is more general in that it would still apply if (e.g.) the square peg were made of glass, and in that respect better. One form of insensitivity that is particularly important when evaluating proposed solutions to Russell's problem is insensitivity to the details of the physics of our world. In particular, the more such an explanation is compatible with various plausible fundamental physical theories of our world, the better it is. In other words, we should privilege those explanations that do not make facts about effective strategies dependent on which of our best candidates for the status of fundamental theory is actually true. One justification for this requirement is that since we do not yet know which of these candidates is correct, an explanation which makes facts about effective strategies dependent on the truth of one of these theories is less epistemically secure than one that doesn't. For instance, since we do not yet know if the dynamics of our world is fundamentally deterministic or indeterministic, it is better to have an explanation of causal phenomena that is compatible with both possibilities. This is one reason not to give much weight to the idea that one may explain the time-asymmetry of causation (hence of strategies) in terms of the temporal asymmetry of GRW's laws of nature.[25] A second justification for this requirement is that if causal phenomena

[25] Cf. the end of §1.2.2.

are dependent on some fine physical details of how our world actually works, there is a question of how we are able to know that there is actually causation in our world. As Kutach (2013, 6) notes, this is a problem for accounts on which causation requires the transfer of a conserved quantity from cause to effect (e.g. Dowe (2000)). Since we are not in a position to tell if in our world quantities like energy or momentum are perfectly conserved or just very nearly conserved, the theory makes it hard to see how we could have evidence for the existence of causal relations in our world. It is therefore better to have an account of causation that doesn't rely on the precise form of laws of nature, such as whether they involve perfectly or approximately conserved quantities.

These considerations on generality shed light on the remark I made earlier that explaining what causation is in our world may involve considering close-by alternative worlds as well. Since a proper explanation of effective strategies shouldn't depend too much on the physical details of our world, checking whether a proposed completion of **(S)** is a good explanation will involve checking whether it holds in a range of possible worlds. For instance, if it turns out that according to the explanation under consideration there is no causation in worlds in which (say) Bohmian mechanics is true and are otherwise just like our world in all easily observable respects, this will count against it. Likewise, since there are worlds that are very much like ours macroscopically but obey Newtonian mechanics, it will count against the theory if it entails that in such worlds there is no causation, or that causation in those worlds is very much unlike causation in the actual world.

A second important explanatory virtue is *unification*. We expect good explanations to exhibit interesting relationships or connections between various phenomena.[26] For instance, what makes Newton's theory a good explanation of motion is in part the fact that it provides a unified account of terrestrial and celestial motion. Here are two respects in which unificatory power plays a role in evaluating solutions to Russell's problem. First, the fact that unification is an explanatory virtue allows us to answer an objection against my proposed criterion of evaluation. The objection is that since it

[26] Whether explanation is *fundamentally* a matter of unification, as Friedman (1974) and Kitcher (1989) argue, is a different question.

focuses solely on effective strategies, the criterion is too anthropomorphic. For instance, it may seem that for all I have said so far, a proposed completion of **(S)** could impose the requirement that only human actions can be causes and still be successful by my lights. After all, since strategies are human actions such a theory may well be able to capture all the facts about effective strategies. But surely there is something defective with any theory that restricts causes to human actions; it is obvious that events that are not actions can be causes. My answer to this objection is that such a theory wouldn't in fact count as a good explanation of effective strategies, as it would lack a crucial explanatory virtue of *unification*. To see this, contrast the theory under consideration with a theory that allows causes to be events other than human actions. The latter is more unificatory in the sense that it makes the pattern that holds between an action and its effects a special case of a wider kind of pattern that can hold for non-actions as well. If both theories are otherwise on a par, the latter, more unificatory theory should be preferred. So on the criterion of evaluation I propose there is a theoretical pressure in favor of theories that do not make causation an anthropomorphic phenomenon, but instead treat causal relations involving human actions as a particular instance of a larger kind of physical pattern.

A second respect in which unification will be important in what follows pertains specifically to the *time-asymmetry* of effective strategies. As we have seen, a proper solution to Russell's problem should explain where in physics it comes from. Now, there are other physical time-asymmetries besides the time-asymmetry of effective strategies. In particular, there is the asymmetry of entropy increase encoded in the second law of thermodynamics and the asymmetry of radiation. One way in which an explanation of the time-asymmetry of effective strategies could (*ceteris paribus*) be better than another is by unifying the time-asymmetry of effective strategies with those other asymmetries, i.e. by making it plausible to think that all these asymmetries derive from the same facts.

A third important explanatory virtue is *reducing puzzlement*. An explanation can earn points by showing how a *prima facie* mysterious phenomenon can be derived from

facts that are not themselves mysterious. For instance, the value of Hamilton's explanation of biological altruism derives in part from its demonstration that a surprising phenomenon (the existence of altruism in a world driven by competition between living organisms) can be accounted for in terms of the principles of natural selection, which are not themselves mysterious (Hamilton, 1964). This virtue will play a particularly important role in what follows, for two reasons. First, note that Russell's arguments leave us with two puzzles regarding effective strategies. It is natural to think that what we can and can't influence is determined at least in part by the laws of nature driving the development of our world. But in light of Russell's argument it is somewhat mysterious why we can influence anything at all since the laws of nature relate global states of the world only; and it is also mysterious why we can influence the future but not the past given the temporal symmetry of laws of nature. (As we will see the second mystery is the most serious, least easily dispelled one.) One of the leitmotivs of the next chapter will be that existing attempts to solve the problem of causation in a physical world have difficulties dispelling these mysteries about effective strategies. As we will see, the problem often takes the following form. A theory of causation identifies causation with some quite complex physical relation R, such that it is *prima facie* not at all obvious that R should underlie effective strategies. In that case, the theory must explain why R in fact underlies effective strategies. But theories of causation often fail to do so. In particular, they often leave it unclear why their favorite physical relation R should be the relation that we should care about when we try to assess the consequences of our actions on desired outcomes, *rather than some close-by relation* R^*. I will close this the problem of *close-by alternatives*. We will see an instance of this problem when we consider Hume's explanation of the difference between causes and effects in §4.2.3.

4.2.3 An Illustration: The Temporal Theory of Causal Direction

To illustrate how these various explanatory virtues can bear on proposed solutions to Russell's problem, I will consider a simple and *prima facie* attractive solution to the

problem of the causal direction and time-asymmetry.[27] This solution relies on the idea that the direction of causation can be explained in terms of the temporal asymmetry itself: what distinguishes causes from effects is precisely that the former come before the latter. Call this the *temporal theory of causal direction*.

There are two markedly different ways to interpret this idea. First, there is a *conventionalist* version. On this view there is no substantive distinction between cause and effect, rather the causal asymmetry is a matter of convention or stipulation. We distinguish causally connected events by calling the earlier one 'cause' and the later one 'effect', but nothing of significance hangs on this. We may call this the *Humean theory of causal direction*, since this is the standard interpretation of Hume's view in the *Treatise*. According to Hume, causation is fundamentally a symmetric relation of constant conjunction and spatial contiguity, on which we conventionally impose an asymmetry in the image of temporal ordering.

Second, there is a *substantive* version of the temporal theory, on which there is a real, non-arbitrary difference between causes and effects. This view presupposes that there is something substantive to the temporal arrow itself - i.e. that the difference between past and future is not a mere matter of stipulation. So the substantive view can be declined in two ways, corresponding to the two substantive views of the temporal arrow. First, there is the view on which there is an intrinsic asymmetry built into the fabric of time itself, so that the past-future direction is metaphysically privileged. Here one finds three-dimensionalist views on which the past exists but the future doesn't, and views on which the past is fixed while the future is open. Another version is Maudlin's (2007) view. Maudlin endorses four-dimensionalism but argues that there is something metaphysically privileged about the past-future direction: earlier states of the world *generate* or *produce* future states, 'production' being a primitive notion.[28] An alternative substantive view of the temporal asymmetry holds that there is no intrinsic

[27] This is only a partial solution to Russell's problem since it says nothing about the problem of localized causes.

[28] Since 'production' sounds a lot like causation, one may think that on Maudlin's view there are primitive causal relations built into the fundamental physical structure of our world. But Maudlinian production lacks one of the central features of causation, namely **Localization**. On Maudlin's view earlier states produce future states via the physical laws. And as we have seen these laws take only global states of the world as inputs.

direction built into the fabric of time, but that the asymmetry between past and future is due to the way in which material contents are distributed along the temporal dimension of the manifold. So on this view the past-future direction is a lot like the up-down direction. There is no intrinsic asymmetry built into the up-down dimension, rather the asymmetry is due to an asymmetry in the way in which physical stuff is distributed along the dimension (here an asymmetry in the distribution of gravitational potential). Perhaps the most famous version of this view is Boltzmann's (1897) view on which the temporal arrow is due to the distribution of *entropy* along the temporal dimension. What gives time its direction is the fact that entropy is lower in the direction that we call the past and higher in the direction we call the future.

All versions of the temporal theory of the causal direction have certain advantages. They have no difficulties explaining why the causal arrow is aligned with the temporal arrow since they identify the two in our world. Also, they offer a straightforward explanation of how we can distinguish which one of a pair of causally connected events is the cause and which one is the effect. But all versions of it to display some of the explanatory virtues described above.

One argument often raised against the temporal theory is that backward causation seems conceptually possible. Thus Lewis writes that

> Careful readers have thought they could make sense of stories of time travel...; hard-headed psychical researchers have believed in precognition; speculative physicists have given serious consideration to tachyons, advanced potentials, and cosmological models with closed timelike curves...It will not do to declare [these phenomena] impossible *a priori*. (1979, 464)

However this isn't a powerful objection against the temporal theory conceived as a solution to Russell's problem, since such a solution need only apply to the actual world. But there is a closely related objection, namely that for all we know our world may allow circumstances in which we can influence the past. In particular, time travel isn't obviously incompatible with the physics of our world, as for all we know the future of our universe might contain closed timelike curves. If so there may be (admittedly exceptional) circumstances in which we may one day be able to influence the past, e.g. by sending a particle back in time. But the temporal theory prohibits the existence of

backward influence in our world.[29] So the theory runs the risk of making the connection between causal and temporal arrows, and thereby to be extensionally inadequate.

Even if there are no closed time-like curves or other circumstances that permit backward influence in our world, their physical possibility means that the temporal theory of causal direction is not as *general* as one might like. One may wish for a theory of the causal direction that is insensitive to the physically contingent fact that there are no closed time-like curves in our world, and thus doesn't rule out the possibility of backward influence (while still entailing that at least in our world the direction of effective strategies is past to future except perhaps in exceptional circumstances). Such a theory would have the advantage of being applicable to a range of physically possible ways the world might be.

Turning to unification, one disadvantage of the temporal theory (at least on some versions of it) is that it fails to unify the causal time-asymmetry with other physical asymmetries. Our best explanation of the asymmetries of entropy and radiation appeal to the laws of nature and boundary conditions of the universe to explain why entropy increases toward the future but not the past and why radiation is always retarded.[30] On many versions of the temporal theory, these facts play no role in making causes precede their effects.[31] The only exception is the variant of the view on which the direction of time is fixed by the direction of entropy. This theory does manage to unify the asymmetry of entropy and the causal asymmetry, by making the latter arise from the former.

Finally, the temporal theory does very little to dissolve the mystery of why despite the symmetry of physical laws we can nevertheless advance our goals in the future direction only. The problem arises most starkly for the Humean, conventionalist version of the view (Dummett, 1954; Price and Weslake, 2009). If the causal asymmetry is a

[29] The temporal theory also entails that EPR-correlations do not involve backward causation. If the goal is to explain effective strategies this is less problematic, as EPR correlations are not exploitable by human agents.

[30] I will summarize the standard explanation of the asymmetry of entropy increase in chapter 2, §2.1.

[31] (2007, 131-135), however, does argue that the temporal asymmetry conceived as an asymmetry of production does contribute to explaining the entropy gradient, by explaining away the atypicality of microstates that lead to lower entropy toward the past. If this is correct, then the facts which on Maudlin's view explain the causal time-asymmetry also contribute to explaining the thermodynamic asymmetry. But see Loewer (2012) for a criticism of Maudlin's argument.

matter of convention and there is no significant difference between cause and effect, why are effective strategies aligned in the same direction as the causal arrow? A mere *conventional* distinction between earlier and later terms of a fundamentally symmetric relation cannot explain why causal relations are exploitable in one direction only. Dummett puts the point as follows:

> If we can observe that an event of a certain kind is a sufficient condition of an earlier event of some other kind, it does not seem to matter much whether we choose to *call* the later event the 'cause' of the earlier or not: the question rather is why we should not use this observed regularity as we use those that operate from earlier to later; why, when we do not know whether or not the earlier event has occurred, we should not bring about the later event in order to ensure that the earlier had occurred. (1954, 28)

This is an instance of the problem of close-by alternatives. Hume's view has difficulties explaining why the relation with which it identifies causation (constant conjunction + temporal precedence) should underlie effective strategies, rather than the close-by (and simpler) relation of constant conjunction.

The problem is less stark if one adopts a substantive account of the temporal asymmetry, although it doesn't disappear entirely. Consider views on which there is an intrinsic temporal asymmetry first. There one might insist that this intrinsic asymmetry is of the right kind to ground an asymmetry of effective strategies. Perhaps we can influence the future but not the past because the future is open whereas the past is fixed, or because influence requires production (in Maudlin's sense). But the concepts of openness/fixity and production are themselves somewhat mysterious, in part because they are primitively modal notions. Consider next the view on which the temporal asymmetry is an asymmetry in material content (e.g. entropy distribution) along the temporal axis. Without a further story as to why the asymmetry of entropy increase bears on the direction of effective strategies, this view still leaves the time-asymmetry of goal advancement mysterious.

4.2.4 Effective Strategies, Practical Rationality and Causal Dependence

To close the explication of my proposed criterion of evaluation, it will be useful to say a bit more about the notion of effective strategies and its connection to causation. In

particular this will help us clarify what *kind* of causal relation should play a role in a proper physical explanation of effective strategies.

I propose to unpack the notion of an effective strategy by looking at its connections with *practical rationality*. The thought here is that information about effective strategies is the sort of information we need to decide where our best interest lies in a decision situation. If I have the option of doing an action a and I know that a is an effective strategy for a desired outcome o, this gives me a *pro tanto* reason to do a. We can spell out this idea more precisely via the following principle, which may be regarded as a sort of implicit definition (or regimentation) of the notion of an effective strategy:

> Suppose that an agent has a choice between doing an action a or not doing a. Suppose moreover, that the agent's only goal is to have a desired outcome o occur. (It is assumed that the agent doesn't know at the time of deliberation whether e occurs.) Then if the agent knows that a is an effective strategy for o, she is rationally required to do a.

For instance, suppose that Suzy has a choice between throwing and not throwing a rock at a window, and that her sole goal is to have the window break. Then if she knows that her throwing a rock is an effective strategy to shatter the window she is rationally required to throw. The requirement that the agent's only goal is to have o occur is crucial. If the agent has other goals, a may have a negative influence on those goals large enough to make a inadvisable after all. Even if Suzy knows that throwing the rock is an effective strategy to break the window, she may not be rationally required to throw the rock if doing so would have other consequences that she values negatively.

Unpacking the idea of an effective strategy in terms of practical rationality has three beneficial consequences. First, in later chapters it will allow us to make use of the powerful and precise formal tools of decision theory to make progress in finding a satisfactory solution to Russell's problem. The second consequence brings us back to the question of the role of intuitions. Earlier on I said that intuitions about effective strategies have a role to play in delimiting the range of phenomena that a proper theory of causation in the physical world should explain. Given the way I have just unpacked the notion of effective strategies, it turns out that the relevant intuitions

are intuitions about what choice is the rational one in a certain decision situation. So the intuitions that matter here are intuitions about practical rationality, not intuitions about causation *per se*. This is an important difference. One issue for the project of conceptually analyzing causation is that in quite a few cases causal intuitions are disputed. Hitchcock (2003) provides a whole range of examples in which people tend to disagree in their intuitions regarding what causes what. The problem here is that it is not clear which of those intuitions (if any) a proper conceptual analysis of causation should accommodate. By contrast intuitions about practical rationality are remarkably stable and widely shared. In the decision-theoretic literature, one finds very few cases in which intuitions about what it is practically rational to do tend to conflict. (Debates in decision theory tend to concentrate on the correct formulation of the general principle of rational choice.) Indeed as we will see the most discussed cases in decision theory - so-called *medical Newcomb problems*[32] - are cases in which it is *very clear* what one should rationally do.[33]

The third consequence is that we are now in a position to say exactly what kind of causal relation should be the target of a proper solution to Russell's problem. The philosophical literature on causation has tended to focus exclusively on what Pearl (2009) calls 'actual causation'.[34] Actual causation is the sort of relation reported in claims like 'The cat caused the vase to break' or 'This solar flare caused an electric disturbance'. It is usually expressed in the past tense, and is the sort of relation that

[32] This family of cases will be discussed extensively in ch. 3.

[33] There is one famous case - the original Newcomb problem (Nozick, 1969) - in which people tend to have conflicting intuitions about what to do. In Newcomb's problem an agent must choose between taking both an opaque box and a transparent box, or taking an opaque box only. The transparent box contains one thousand dollars. The opaque box contains either a million dollar or nothing, depending on the prediction made by a demon. If the demon predicted that the agent would take the opaque box only, he put a million dollars in it. If he predicted that the agent would take two boxes, he put nothing in the opaque box. The demon is very reliable at such predictions, and the agent knows this. Some people have the intuition that the agent should take one box only (since if she does so she is likely to get the million). Others insist that she should take the two boxes. Since the prediction has already been made, the agent's choice has no influence on the contents of the boxes, so that if she were to choose the opaque box only she would be a thousand dollars less rich than if she were to take the two boxes. As we will see, medical Newcomb problems have a structure similar to Newcomb's problem but are much more realistic. Moreover in medical Newcomb problems there is no debate about which action is the correct one.

[34] Actual causation is also often called 'singular causation', 'token causation' or simply 'causation'. As we will see these names are misleading since actual causation is not the only causal relation, and not the only one that holds between singular events.

matters for assessing moral and legal responsibility. For our purposes the crucial point is that actual causation is not the relation that matters for decision-making, as Hitchcock (2013) clearly demonstrates.[35] He contrasts the following two cases:

> *Suzy.* Suzy has a choice between throwing a rock at a window or not doing so. Her sole goal is to have the window shatter. She knows that if she throws the rock, the window will shatter as a result, and that if she doesn't throw the rock, the window will remain intact.
>
> *Preempting Suzy.* Suzy has a choice between throwing a rock at a window or not doing so. Her sole goal is to have the window shatter. She knows that, independently, Billy will throw a brick at the window. She knows that if she throws her rock, it will strike the window a few seconds before Billy's brick arrives at the window's location. She also knows that if she doesn't throw her rock, Billy's brick will strike the window.

In both situations, if Suzy throws her rock her action will be an actual cause of the window breaking. (It would be appropriate to say after the fact: 'Suzy throwing the rock caused the window to break'.) If actual causation were the sort of relation that matters for decision-making, in both cases Suzy should rationally throw the rock. But whereas this is true in the first case, this isn't true in *Preempting Suzy.* Intuitively, in the latter case, Suzy should be indifferent between throwing and not throwing, as the outcome she desires will occur whatever she does. If the causal relation that matters for decision-making is not actual causation, what is it instead? *Preempting Suzy* provides an answer. Clearly there Suzy is not rationally required to throw the rock because, by contrast to *Suzy*, her doing so doesn't *make a difference* to the outcome she cares about. As we might also put it, whether or not the desired outcome occurs doesn't *causally depend* on her throwing the rock. So the relation that matters for decision-making is *difference-making* or *causal dependence* (I will use the two phrases interchangeably).[36] Correspondingly, the target of a proper solution to Russell's problem should be causal dependence, not actual causation. To explain effective strategies, one need only explain what the former relation is. By way of making the notion of causal dependence clear, I note that *counterfactual* dependence is a good test for it (Lewis, 1973a). When e

[35] See also Hall (2004, 268-70).

[36] The phrase 'difference-making' comes from Lewis (1973a). The phrase 'causal dependence' is widely used by decision theorists (e.g. Joyce (1999)) to designate the sort of causal relation relevant for practical purposes.

causally depends on c, the following two counterfactuals (or some very similar ones) are true:

(1) If c were to happen, e would happen.
(2) If c were not to happen, e would not happen.

The fact that causal dependence and not actual causation is the target here is important for the following reason. The cases that have turned out to be the most difficult to handle for conceptual analysts (whose main focus is on actual causation) are cases in which the actual cause is not a difference-maker. I have in mind familiar cases of preemption and overdetermination.[37] Consider *Preempting Suzy* again. This is a case of *late preemption*, where one process (Suzy's rock breaking the window) prevents another (Billy's brick striking the window) to go to completion. Suppose, alternatively, that Suzy throws the rock and breaks the window. Billy would have thrown his brick just in case Suzy hadn't thrown her rock. This is a case of *early preemption*. Finally, suppose that both Suzy and Billy throw their rocks; both rocks strike the window at the same time, with sufficient momentum to break it. This is a case of *overdetermination*. In all those cases Suzy's throwing her rock is intuitively an actual cause of the window breaking, but not a difference-maker for it. Conceptual analyses of actual causation have trouble handling those cases or more refined varieties of it, and this is one of the prominent reasons to be skeptical about the feasibility of conceptual analysis.[38] But this is no reason to be skeptical about the feasibility of solving Russell's problem, since such a solution should be a physical theory of *causal dependence* only. It need not accomplish the difficult task of explaining in physical terms the difference between preempting and preempted causes to be successful.

Two important remarks before I close this section. First, one worry with focusing exclusively on difference-making and leaving actual causation aside is that Russell's arguments also raise a challenge for the existence of actual causation. After all, actual

[37] See also cases of trumping (Schaffer, 2000b).

[38] For an in-depth study of the difficulties raised by these cases and variations thereof for conceptual analyses of actual causation, see Paul and Hall (2013). Process theories of actual causation (e.g. Dowe (2000)) handle these cases beautifully but fail in others, such as cases of double prevention (Schaffer, 2000a).

causation is itself a paradigmatically local, asymmetric relation, so that in light of Russell's arguments it is not clear how physics can leave space for actual causation. But positing actual causation is indispensable since the notion plays a crucial role in our moral and legal practices. Thus one may worry that focusing exclusively on difference-making will not allow us to solve all the problems raised by Russell. To this I respond, first, that providing a good explanation of how the kind of causal relation that matters for decision-making can arise from physics would already go a very long way toward alleviating the worries raised by Russell. Second, it is rather uncontroversial that there are deep conceptual links between difference-making and actual causation. In light of this, it is not unreasonable to think that actual causation inherits its locality and asymmetry from difference-making, so that a good physicalist explanation of the latter would thereby solve the problem of the compatibility of actual causation with fundamental physics.

The second remark pertains to the relations between difference-making and counterfactuals. As I noted earlier, difference-making is closely related to counterfactual dependence. For instance, it is natural to say that in *Suzy*, throwing the rock makes a difference to the window breaking because were Suzy to throw the rock the window would break, whereas if she were not to throw the rock the window would remain intact. Thus, it may seem that a theory of difference-making will necessarily be what is called in the literature a *counterfactual theory*. (Lewis, of course, is the patriarch of counterfactual theorists.) However, it is important to note that counterfactual theorists are committed to two controversial assumptions. The first is that the semantics for statements like (1) and (2) above differs starkly from the semantics of indicative conditionals. This assumption can be cashed out as follows. There is a stark difference in meaning between the following two past-tense conditionals:

(3) If Oswald didn't shoot Kennedy someone else did.
(4) If Oswald hadn't shot Kennedy someone else would have.

To see this, note that if you don't believe there was a backup shooter, you will accept (3) but not (4). Counterfactual theorists assume that the right semantics for (1) and

(2) is much more similar to the semantics of statements like (3) than the semantics of statements like (4). One need not accept this assumption at the outset. It might be that the right truth-conditions for conditionals that matter for decision-making (such as (1) and (2)) look more like the semantics for (4) than the semantics for (3).[39] The second assumption made by counterfactual theorists is that causal dependence just is counterfactual dependence. One might hold instead that even if causal dependence can be usefully captured by ordinary language statements like (1) and (2), the most fruitful physical explanation of causal dependence should make use of devices (such as conditional probabilities) that do not obey the logic of ordinary language counterfactuals. In these two respects, a successful theory of causal dependence need not be a counterfactual theory.

4.3 Causal Dependence: Valence, Relata and Adicity

I will close this chapter with some formal remarks about causal dependence as I understand it, thereby bringing to the foreground some implicit assumptions I made throughout the chapter.

Note that causal dependence can be either *positive* or *negative*. For instance, Suzy throwing the rock is a positive difference-maker for the window breaking, whereas Johnny taking anti-fever medication is a negative difference-maker for him having a fever later. Throughout the later chapters I focus almost exclusively on positive causal dependence, so that 'c causes e' should be read as 'c is a positive difference-maker for e'. It should be very straightforward to extend what I will say about positive causal dependence to the case of negative causal dependence.

I take the relata of causal dependence to be *singular events*. Thus causal dependence as I understand it is a relation of singular causation. It is often assumed in the literature on causation that there is a meaningful distinction to draw between two kinds of causal relations, *general* causation and *singular* causation (e.g. Sober (1985); Eells (1991)). General causation is supposed to be a *sui generis* relation holding between properties or event types, and expressed in statements such as 'Smoking causes cancer' or 'Solar flares

[39] This position is defended by e.g. DeRose (2010).

cause electric disturbances'. However, this position seems to multiply causal relations beyond necessity. One may instead regard general causal statements as quantified statements over instances of singular causation Lewis (1973a); Carroll (1991). The relevant quantifier is probably the *generic* quantifier (Eagle, 2014), so that 'Smoking causes cancer' means something like 'Normal or stereotypical episodes of smoking cause episodes of lung cancer', where the relevant relation of singular causation is *actual* causation. Insofar as a theory of the place of causal dependence in our physical world can also help explain how relations of actual causation fit in our physical world, it can thus also help explain how the fundamental physics of our world leaves space for true general causal statements. Thus, given our purposes here, it is appropriate to focus entirely on the (singular) relation of causal dependence.

The assumption that the relata of causal dependence are (singular) *events* is more a matter of convenience than a deep metaphysical commitment. There are important questions about whether singular causal relata really are events, or if we should take them to be particular facts (Mellor, 1995), states of affairs (Armstrong, 1997) or perhaps aspects (Paul, 2000). But these questions are orthogonal to the issues arising from Russell's attack on causation, and I will ignore them in what follows. As far as I can see, although it is couched in terms of events the theory of causal dependence I will propose is compatible with most existing views about the nature of singular causal relata.

I assume, as is standard in the literature, that an effect and its cause must be *distinct* events in the sense of Lewis (1986c). Distinctness means that the events are not identical, nor is one a part of the other, nor does one metaphysically or conceptually implies the other. This requirement is essential to exclude dependencies between events that are logical or metaphysical rather than properly causal. Without this requirement, all the theories I will discuss would count the event of Larry laughing loudly as a cause of Larry laughing. Clearly the relation between these events is not causal, because the two events are not distinct.

Finally, a word about the adicity of causal dependence. So far I have talked as if causal dependence were a binary relation. However, it is more appropriate to regard

it as a contrastive four-place relation of the form: *c rather than c* makes a difference to e occurring rather than e**. We may call *c** the *cause-contrast* and *c** the *effect-contrast.*[40] One argument for contrastivity on the cause side is that binary causal dependence is sometimes ill-defined (Hitchcock, 1996b). Does Jane smoking one pack a day makes a difference to her getting lung cancer? The answer seems to depend on what one contrasts the putative cause with. Her smoking a pack a day rather than not smoking at all made a difference, but her smoking one rather than two packs a day didn't. Once one accepts contrastivity on the cause side, one needs to accept it on the effect side as well to make sense of causal chains. As Schaffer (2014) puts it: 'In a causal chain the effect at the first link serves as the cause at the second. For this to be possible, cause and effect must be formally exchangeable: the same structure must flank both sides of the relation' (2014, §1.3). So if there is a cause-contrast, there must be an effect-contrast as well. For simplicity, in what follows I will often leave the contrasts implicit and write 'c causes e' instead of the more cumbersome 'c rather than c^* makes a difference to e occurring rather than e^*'; and I will use the negative descriptions '$\sim c$' and '$\sim e$' to designate the cause- and effect-contrasts. In the particular examples I will use, the intended contrasts should be clear from the context.

[40] Hitchcock (1996b); Maslen (2004); Schaffer (2005) defend contrastivism about actual causation. Their arguments easily transfer to causal dependence.

Chapter 2

Current Solutions to Russell's Problem: A Critical Examination

In this chapter I will review current solutions to Russell's problem, using the methodological and evaluative guidelines laid out in the previous chapter. Many of the authors I will discuss do not explicitly present their theories as responses to Russell's problem. Rather, their goal is to conceptually analyze causation. In accordance with the guidelines of ch. 1, I won't try to evaluate whether their theories succeed as conceptual analyses. Rather, I will examine whether they provide a good physical explanation of causation in our world that illuminates effective strategies. Note also that many of the authors I will discuss are chiefly concerned with actual causation. This isn't to say that they do not talk about difference-making. Many philosophers of causation follow Lewis (1973a) in thinking that difference-making is the right starting point for explaining actual causation. Accordingly, these authors offer detailed theories of difference-making. Most of this chapter will be devoted to examining the prospects of the two main approaches to difference-making.

The first one I will call the *Lewisian approach*, since its guiding idea was first articulated by Lewis (1973a, 1979). Its guiding idea is that difference-making should be understood as follows. Take the situation (or 'world') closest to actuality in which c happens, and the situation closest to actuality in which c does *not* happen. If e

happens in the former situation but not in the latter, then c makes a difference to e. Although Lewis's version of this idea faces strong objections, a much stronger account along Lewisian lines has been offered by Albert (2000), Kutach (2002, 2007) and Loewer (2007, 2012). I examine Lewis's approach in §1 and the Albert-Kutach-Loewer account in §2.

I call the second approach the *probabilistic approach*. It has been endorsed by Reichenbach (1956), Papineau (1985, 1993, 2001b), Spohn (2001) and Field (2003), among others. This name is somewhat misleading insofar as it suggests that what makes this approach different from the Lewisian approach is the appeal to probability. True, proponents of the probabilistic approach all agree that even if we assume determinism we need non-trivial conditional probabilities to explain difference-making - a thesis that Lewis rejected.[1] But as we will see, the Lewisian account developed by Albert, Kutach and Loewer also gives a central role to conditional probability. Rather, what makes the probabilistic approach distinctive is that c being a difference-maker for e is determined by c's and e's respective places in a *probabilistic web*. To be a cause is to occupy a certain position in a probabilistic network involving the cause, the effect, and other events. I examine the probabilistic approach in §3.

I will argue that current versions of both the Lewisian and the probabilistic approach all fail to provide a satisfactory solution to Russell's challenge. Nevertheless, our review of these two approaches will also allow us to extract important ideas about causal dependence, which I will later use in my own account. Indeed, the account I will present in chapter 4 owes much both to the Lewisian and to the probabilistic approach.

There is one prominent solution to Russell's problem that I won't discuss in this chapter, namely the one offered by Price.[2] Price endorses an *agency* theory of causation, on which what makes e causally dependent on c is the fact that an agent who can choose to do c can thereby influence the occurrence of e. Most agency theories are explicitly non-reductive (or clearly fail at reducing causation); Price's view is the only sustained attempt to turn this idea into a reductive account of causation that can explain the

[1] See Postscript B in Lewis (1986b).
[2] See Price (1996, 2007, 2012), as well as Price and Weslake (2009).

compatibility of localized asymmetric causation with fundamental physics. The reason I won't discuss it here is that my account relies on the same guiding ideas as Price's, although it develops them in a very different way. Thus it is more appropriate to discuss Price's views in chapters 3 and 4, when I expose my solution to Russell's problem.[3]

1 Lewis's Theory of Causal Dependence

Lewis's theory is a *counterfactual* theory in the sense defined in ch. 1, §4.2.3. That is, on Lewis's view c makes a difference to e just in case the following two statements are true:

(1) If c were to happen, e would happen.
(2) If c were not to happen, e would not happen.

where (1) and (2) are ordinary language counterfactuals. This works only for determinism. In the indeterministic case, Lewis offers a somewhat different counterfactual theory.[4] To keep the discussion manageable I will focus entirely on the part of Lewis's

[3] In addition to the Lewisian, probabilistic and Pricean approaches to causation, there is another approach to causation with explicit reductionist ambitions, the *causal process* approach. Its most detailed and promising development is due to Dowe (2000). On Dowe's view, very roughly, a causal connection between a cause and its effect involves the exchange of a conserved quantity (such as energy or momentum) between the cause and the effect. For instance, what makes the cue ball rolling and the 8 ball sinking causally connected is that when the cue ball and the 8 ball collide, they exchange energy and momentum.

I won't discuss this approach any further in this dissertation, for several reasons. The first and main one is that causal process theories, whose focus is on actual causation, do not offer a theory of causal *dependence*. To see this, consider again *Suzy* and *Preempting Suzy* from chapter 1. In both cases, when Suzy's rock strikes the window there is a transfer of conserved quantities between the rock and the window, so that there is a causal connection between them according to the process theory. So the process theory captures the similarities between the two cases. But here we are interested in the *differences* between the two cases: we are interested in explaining what makes Suzy's throw a difference-maker in the former case but not in the latter. Another reason is that as Dowe himself points out the causal process theory by itself doesn't capture the asymmetry of causation: it only captures causal connection. (The notion of an exchange of conserved quantity is symmetric.) Indeed, to explain the asymmetry of causation Dowe (2000, ch. 8) relies on Reichenbach's theory of the fork asymmetry, which will be discussed later in this chapter. A last reason is that the causal process theory seems ill-placed to explain the practical relevance of causation. It is not at all obvious why we should care about exchanges of conserved quantity between actions and desired outcomes in the context of rational decision-making.

[4] Suppose that Suzy throws the rock and breaks the window. If she hadn't thrown the rock, the indeterministic laws of nature say that the window would have had a small chance of shattering spontaneously. Here intuitively Suzy makes a difference to the window breaking, but the counterfactual 'If Suzy hadn't thrown the rock, the window wouldn't have broken' is false: the window *might* have

theory that deals with the deterministic case.[5]

1.1 Lewis's Semantics for Counterfactuals

The intuitive idea that drives Lewis's semantics for counterfactuals is that the statement 'if c were to happen, e would happen' is true iff in a situation in which c happens and which otherwise diverges from actuality as little as possible, e also happens. To cash out this idea, Lewis (1973b) uses the notion of a *comparative similarity* relation between worlds. A possible world W is said to be more similar (or closer) to the actual world than another possible world W' just in case W resembles the actual world more than W' does.[6] Let's say that a possible world in which c happens is an c-world. We can then state the Lewisian truth-conditions for counterfactuals as follows:

> **Comparative Similarity Analysis (CSA)**. 'If c were to happen, e would happen' is true in the actual world just in case either there are no c-worlds, or some c-world where e happens is closer to the actual world than any c-world where e doesn't happen.

Lewis stipulates that the actual world is closest to itself. From this and **CSA**, it follows that when c and e are actual events, the counterfactual 'if c were to happen, e would happen' is automatically true. Thus on Lewis's view, when c and e are actual, c makes a difference to e just in case the counterfactual 'if c were not to happen, e wouldn't happen' is true.

To be made sufficiently precise **CSA** needs to be supplemented with an account of the relation of comparative similarity. Crucially, for Lewis the comparative similarity

broken. What is true, however, is that if Suzy hadn't thrown the rock, the window's chance of breaking would have been much smaller than it actually was. Thus in the indeterministic case Lewis says that c makes a difference to e just in case the following two counterfactuals are true:

- If c were to happen, e's chance of occurring would be x
- If c were not to happen, e's chance of occurring would be y

where x is higher than y.

[5] As we will see this will be enough to show that Lewis's approach as a whole suffers from severe problems.

[6] Lewis famously espouses the view that possible worlds are real concrete entities ontologically on a par with the actual world. But one can endorse the Lewisian framework without committing oneself to this extreme modal realism. His semantics is compatible with less outlandish views about possible worlds (e.g. the view that possible worlds are maximal consistent sets of propositions).

relation that matters for counterfactual evaluation is *not* our intuitive notion of overall similarity between worlds. His reasons for saying so trace back to a problem for **CSA** pointed out by Fine (1975), the *future similarity objection*. Take the counterfactual 'if Nixon had pushed the button on December 31, 1971, there would have been a nuclear war'. This counterfactual is intuitively true, but (Fine contends) on **CSA** it comes out as false. Intuitively, a world in which Nixon pushes the button but no nuclear war ensues (thanks to a failure of the missile launching system, perhaps) is overall more similar to the actual world than a world in which Nixon pushes the button and a nuclear holocaust happens. After all, in the former world post-1971 history looks much more similar to ours than in the latter world. Consequently, **CSA** appears to wrongly count the counterfactual as false.

Lewis's response in his (1979) is to contend that not all respects of similarity matter for counterfactual evaluation. For Lewis, there are two respects of similarity that matter: similarity with respect to laws of nature and similarity with respect to particular matters of fact. A world is more similar to actuality than another the fewer *miracles* (violations of the actual laws of nature) it contains. And a world is more similar to actuality the more it perfectly matches the actual world in particular matters of fact, i.e. instantiations of fundamental physical properties at particular spacetime points. Now, given bi-directional determinism, a world that doesn't violate actual laws of nature and perfectly matches the actual world in particular matters of fact at a certain time must match the actual world in particular matters of fact at all times. A world in which a non-actual antecedent is true must either differ from our world in certain respects at all times or be a world in which the antecedent comes about as the result of a miracle. Hence these two respects of similarity are in tension. Lewis's view allows for a trade-off between them. A world with an extensive region of perfect match to the actual world can be considered very similar to actuality, as long as the match is achieved at the cost of a small, local miracle. More specifically, Lewis (1979, 472) claims that which of two worlds is more similar to the actual world is determined by the following system of priorities:

S1 It is of the first importance to avoid big, widespread, diverse 'miracles' (violations of actual laws of nature).

S2 It is of the second importance to maximize the size of the region in which perfect similarity to the actual world in matters of particular facts prevails.

S3 It is of the third importance to avoid even small, localized miracles.

S4 It is of little to no importance to maximize the size of the region in which approximate similarity to the actual world in matters of particular facts prevails.

Lewis argues that if we plug the similarity relation generated by these standards into **CSA**, the Nixon counterfactual is true.[7] To show this, he considers four worlds in which Nixon pushes the button as candidates for the status of world most similar to the actual world $W_@$. The first world W_1 perfectly matches our world in particular matters of fact up to December 31, 1971, at which point a small miracle occurs that leads Nixon to push the button. (This small miracle might be the firing of certain neurons in Nixon's brain, for instance.) The second world W_2 conforms perfectly to the actual laws at all times. Given bi-directional determinism, since W_2 contains a non-actual event (Nixon pushing the button) it must differ from our world in particular matters of fact at all times. In particular, in W_2 the past is different from the actual world. In W_1 and W_2 a nuclear war occurs. The third world W_3 is a world like W_1, but in which a second small miracle (perhaps a failure of the missile launching system) occurs shortly after Nixon pushes the button so that the history of that world after 1971 looks very similar to ours, although as we will see it isn't an exact match. The last world W_4 is a world like W_1 but in which a second miracle happens after Nixon pushes the button that leads to a post-1971 history that perfectly matches ours in particular matters of fact.

Lewis claims that on [**S1**]-[**S4**], W_1 is more similar to $W_@$ than the three other worlds. First, W_1 is more similar than W_2. The region of perfect match is much greater in W_1, and this perfect match is achieved at the cost of a small miracle only. Likewise, W_1 is more similar to $W_@$ than W_3. Note that at the time of the second miracle, W_3 doesn't perfectly match $W_@$: for instance, at that time, Nixon has a memory of

[7] In fact [**S1**]-[**S4**] only determine a *family* of similarity relations depending on the weight attached to each. I leave this complication aside.

pushing the button in W_3 but not in $W_@$. Since after the second small miracle W_3 evolves lawfully, it follows from determinism that each complete time-slice of W_3 after the second miracle fails to perfectly match the corresponding time-slice of the actual world in particular matters of fact. So W_1 and W_3 are perfectly similar to the actual world up to December 31, 1971; after this time W_3 is more approximately similar to our world, but overall W_1 has one less small miracle than W_3. Since it is more important to avoid even small miracles than to achieve approximate similarity, W_1 is more similar to $W_@$. Finally, Lewis also argues that W_1 is more similar to $W_@$ than W_4. In W_1, shortly after Nixon pushes the button there are many minute respects in which W_1 differs from $W_@$. In the former but not in the latter, Nixon remembers pushing the button; the button is slightly hotter; there are electric signals running along the wires connecting the button to the missiles, and so on. As we might put it, in W_1 but not in $W_@$ there are many *traces* of Nixon having pushed the button. To achieve *perfect* similarity with the actual world after the time at which Nixon pushes the button, W_4 must therefore contain many diverse miracles - each one consisting in the sudden disappearance of a trace of Nixon's action. Since it is more important to avoid diverse miracles than to secure perfect match, W_4 counts as less similar to $W_@$ than W_1. Since in W_1 a nuclear war occurs, Fine's counterfactual comes out as true.

1.2 The Lewisian Solution to Russell's Problem

We can now start evaluating whether Lewis's theory of causal dependence offers a satisfactory solution to Russell's problem. A clear virtue of the account is that it straightforwardly reduces causal dependence to fundamental physical facts. On Lewis's view, causal dependence is determined by the criteria of similarity, which are themselves exclusively sensitive to actual physical laws and distribution of fundamental physical properties across spacetime. Another virtue of the account is that it provides an elegant solution to the localization problem. Let's consider again the case where Suzy throws a rock that breaks a window; we assume that no backup is present so that intuitively Suzy throwing the rock makes a difference to the window breaking. On Lewis's view, what makes it so is that (a) actually, Suzy throws the rock and the window breaks

and (b) if Suzy hadn't thrown the rock the window wouldn't have broken. In turn, the truth of this counterfactual doesn't require Suzy's action to nomologically determine whether the window breaks. What makes the counterfactual true is that in the closest world in which Suzy doesn't throw the rock the window remains intact; and this is compatible with the existence of nomological situations in which Suzy doesn't throw but the window still breaks of the presence of a backup. Since by hypothesis there is no such backup in the actual world, such a world will be further from actuality than a world in which Suzy doesn't throw and there is no backup so that the window doesn't break. So unlike **Determination**, Lewis's view enables localized events to count as causes. But Lewis's view also preserves the close intuitive connection between causation and laws of nature that made **Determination** attractive. Laws of nature come into play in partially determining which possible worlds matter for counterfactual evaluation.

Let's turn to Lewis's solution to the problem of the causal asymmetry. Lewis (1979) argues that the package view constituted by **CSA** and [**S1**]-[**S4**] correctly makes difference-making time-asymmetric in our world. Let's come back to Lewis's solution to the future similarity objection. As we have seen, if Lewis's argument is correct the closest world in which Nixon pushes the button is W_1, a world exactly similar to ours up to shortly *before* Nixon pushes the button, and very different *after*. Thus, on Lewis's view whether or not Nixon pushes the button makes a lot of difference to the *future*, but none or little to the *past*.

This temporal asymmetry isn't built into Lewis's account of counterfactuals. The similarity metric in itself doesn't require counterfactual dependence to go from past to future only, and doesn't explicitly privilege past similarity over future similarity. (A good thing, since otherwise the account would face pretty much the same problems as the Humean theory of the causal direction.) Where, then, does the asymmetry come from? Lewis's answer is that the time-asymmetry of counterfactual dependence is the product of a contingent asymmetry displayed by our world, the *asymmetry of miracles*. The asymmetry of miracles is the following fact. Take any counterfactual with a non-actual antecedent c, and let t be the time of c's occurrence. Then usually it only takes a small miracle for a world exactly similar to $W_@$ up to shortly before t to diverge from

the actual world so as to make c happen. By contrast, it takes a lot of big, diverse miracles for an c-world to be exactly similar to our world at all times shortly after t. Divergence is easy, but convergence is hard. In the case of Fine's counterfactual, this putative asymmetry is what makes it the case that W_1 is more similar to $W_@$ than W_4. More generally, the reason why what Nixon does makes a difference to the future but not the past is that there are worlds with the same history as ours in which Nixon comes to push the button as the result of a small miracle; but there are no worlds in which a small miracle happens that ensures convergence to the actual world after Nixon pushes the button. As Lewis puts it, 'the asymmetry of counterfactual dependence arises because the appropriate standards of similarity, themselves symmetric, respond to this asymmetry' of miracles (1979, 473).

Lewis argues that the time-asymmetry of miracles is the product of another temporal asymmetry of overdetermination. He defines a determinant of an event as 'a minimal set of conditions jointly sufficient, given the laws of nature' for the occurrence of the event (1979, 474). An event is overdetermined when it has two or more determinants. The time-asymmetry of overdetermination is the putative fact that in our world a given localized events has many determinants in its future but few if any in its past. Since the link between a determinant and what it determines is nomological, it takes a miracle (a violation of the laws) to break the link. So if an event has many postdeterminants it will take many miracles to erase those traces of the event and get convergence to the actual world. But if the event only has one or no pre-determinant it will take only a small miracle to get a divergence from the actual world. Thus the asymmetry of overdetermination underwrites the asymmetry of miracles.

To motivate the claim that our world displays this time-asymmetry of overdetermination, Lewis appeals to two examples. The first one is the fact that Nixon pushing the button should leave many traces after the fact (fingerprints on the button, memories in Nixon's brain, and so on) but not before. The second one is Popper's (1956) example of the circularly diverging waves originating from a stone dropped in a smooth pond. In Lewis's words:

> There are processes in which a spherical wave expands outward from a point source to

> infinity. The opposite processes, in which a spherical wave contracts inward from infinity and is absorbed, would obey the laws of nature equally well. But they never occur. A process of either sort exhibits extreme overdetermination in one direction. Countless tiny samples of the wave each determine what happens at the space-time point where the wave is emitted or absorbed. The processes that occur are the ones in which this extreme overdetermination goes toward the past, not those in which it goes toward the future. I suggest that the same is true more generally. (1979, 475)

As this passage shows, on Lewis's view the time-asymmetry of overdetermination is a physically contingent fact. Nothing in the laws of nature dictates that localized events must have few predictors but leave many traces. The time-asymmetry of miracles and hence the time-asymmetry of difference-making themselves inherit this contingence. Lewis also points out that the time-asymmetry of overdetermination may be a local matter only, one that fails to hold in distant regions of the universe. If so, the temporal asymmetry of difference-making may itself be a local matter. Thus Lewis's theory leaves a loophole for backward causal dependence in circumstances where there is no overdetermination of the past by the future. (It thereby escapes one of the problems I mentioned for the temporal theory of the causal direction.[8]

This account can be reformulated in a way that conceptually separates the direction of causation from its time-asymmetry. Say that a world displays an arrow of overdetermination just in case in this world, localized events are overdetermined in some direction of time but not in the other. On Lewis's view, what makes it the case that causes don't depend on their effects (and thus that causation has a direction) is that causes lie upstream of effects with respect to the arrow of overdetermination. The time-asymmetry of causation is the product of the contingent alignment of the arrow of overdetermination with the past-future direction.

1.3 Problems for Lewis's Theory

Lewis's theory offers a clear and detailed explanation of the compatibility of localized asymmetric causation with the globality and symmetry of physical laws. But it also

[8] Lewis doesn't explicitly relate this putative physical asymmetry to the physical asymmetries of entropy and radiation - something we might expect from a proper solution to Russell's problem (see ch. 1, §4.2.2). This isn't a major problem: since Popper's example is a prime example of these three asymmetries one may plausibly hypothesize that there is a unified physical explanation for all of them.

suffers from two severe defects.

The first problem is that Lewis's theory leaves it mysterious why causal dependence as he construes it should be the relation that we should care about in the context of rational decision-making. In particular, it leaves it mysterious why causal dependence *rather than other close-by relations* is the one that matters for the purposes of goal advancement. (In that respect, it encounters a version of the problem of close-by alternatives: see ch. 1, §4.2.2.) The source of the problem is the similarity metric generated by the standards [**S1**]-[**S4**]. As Horwich (1987, 172) notes, these standards are baroque and without pre-theoretical plausibility. In fact, Lewis's sole justification for those standards is that they make true those counterfactuals that we intuitively think are true. If so, this raises the question as to why the relations picked by *those* standards should matter for rational decision-making, rather than some other neighboring relation determined by somewhat different standards. As Woodward puts the point:

> What is the larger point or rationale that lies behind our use of [Lewis's] standards? For example, why don't we employ a set of standards in which [**S3**] is weighted more heavily than [**S2**], or in which, in contrast to [**S4**], some rather than little or no weight is attached to approximate similarity in matters of particular fact? To respond that the standards [**S1**]-[**S4**] are preferred because they are the standards that are reflected in 'our' notion of causation simply invites queries about why that notion is so special. Why should we not (why, in fact, did we not) develop a notion of 'smausation' instead, connected to counterfactuals in the way that 'causation' is in Lewis's theory, but according to which counterfactuals are evaluated by some different set of similarity criteria? (2003, 137)[9]

Note that on the alternative sets of standards mentioned by Woodward, it may very well be the case that the past but not the future counterfactually depends on the present. So insofar as Lewis leaves it mysterious why we shouldn't use those alternative standards, his account has difficulties explaining why we can influence the future but not the past.

To summarize: even if Lewis has managed to identify a physical relation that is both localized and asymmetric (and as we will see there are strong reasons to doubt this), it is unclear why it is *this* relation that we should care about when trying to determine what would be the best strategy to promote the occurrence of a desired outcome.

The second problem concerns Lewis's account of the direction and time-asymmetry of difference-making. The problem is that there are in fact no asymmetries of miracles

[9]With some notational changes for consistency. For similar remarks see Loewer (2007, 323).

and overdetermination in our world. More precisely, Lewis is certainly right that there is an asymmetry of traces (in a pre-theoretical sense of 'traces') in our world. Actual events often leave many traces in their future, but few in their past. But this asymmetry isn't an asymmetry of overdetermination, and it is not of the right kind to underwrite an asymmetry of miracles.[10] In light of the considerations of ch. 1, one can easily see why our world couldn't exhibit an asymmetry of overdetermination. This asymmetry, remember, is the alleged fact that any actual event e has many more localized post-determinants than pre-determinants, where a determinant is a set of conditions that physically determines e. But we saw in chapter 1 that on all our best physical theories of the world, no localized set of conditions is physically sufficient for anything. Nothing less than an entire cross-section of its past or future light-cone can physically determine an event. There are no localized determinants in our world, and *a fortiori* no asymmetry in their temporal distribution.

Formulated in this abstract way, this argument against the asymmetry of overdetermination may leave one puzzled. Take, for instance, the set of traces of Nixon having pushed the button (the fingerprints, Nixon's memories, the signals running along the wires, and so on). The argument of the previous paragraph entails that it is physically possible for those traces to come about without Nixon having pushed the button earlier.[11] This may seem incredible. What would such a physical process look like? One would like to see a concrete explanation of how traces of an event can lawfully appear without the occurrence of the event itself. There is a beautiful argument due to Albert[12] and Elga (2001) that does exactly this.[13] The Albert-Elga argument also offers a direct rebuttal to Lewis's claim that there is an asymmetry of miracles. This is important since one might think that even if there is no asymmetry of overdetermination there might still be an asymmetry of miracles. After all, even if it is physically possible for the traces of Nixon pushing the button to arise without the button being pushed,

[10] As we will see in the next section, there is a more plausible account of the asymmetry of causal dependence that appeals to the asymmetry of traces, but doesn't construe it as an asymmetry of overdetermination.

[11] Note that here I use the word 'trace' in a non-factive sense.

[12] Albert made the point in a seminar at Princeton in 1996 (Loewer, 2007, 313).

[13] See also Frisch (2005, ch. 8).

it may still take many widespread miracles to produce these misleading traces. If so Lewis's claim that worlds that converge onto the actual world's history are very distant would still stand. The Albert-Elga argument shows that, on the contrary, converging worlds need not contain more than a small, localized miracle.

To summarize the argument I will follow Elga's (2001) presentation. Elga asks us to imagine the following situation. At 8:00am Gretta cracks an egg into a hot frying pan, and five minutes later the egg is cooked in the pan. There is no backup egg-cracker in the vicinity, so that the counterfactual 'if Gretta hadn't cracked the egg at 8:00am the egg wouldn't be cooked at 8:05am' is true. At 8:05 there are many traces of Gretta having cracked the egg: her memories, the pieces of egg stuck to the outside of the pan, and so on. Elga shows how to build a world in which Gretta doesn't crack the egg, but a small miracle happening between 8:00 and 8:05 that ensures that at 8:05 things are exactly like in our world. (In particular, all the traces of the egg having been cracked are there.) This shows that there is no asymmetry of miracles: convergence can be ensured by a small miracle. Moreover, it explains concretely how the traces of Gretta cracking the egg could lawfully appear without Gretta having actually cracked the egg.

The procedure to build the relevant world is as follows. For simplicity, let's assume that the laws of our world are those of classical mechanics.[14] Elga's argument exploits the fact that classical mechanics is *time-symmetric*. A theory is time-symmetric when applying a time-reversal operator to a sequence of states allowed by the theory yields a sequence of states that is also allowed by the theory. In classical mechanics, the relevant operator takes a state of the world and outputs another state exactly like the input except that all the velocities of particles are reversed. Now let S_0 be the state of the actual world at 8:00am, and S_1 the state of the actual world at 8:05am. And let Z_1 be the time-reverse of S_1 and Z_0 the time-reverse of S_0, and call the physically possible world in which the Z_1 to Z_0 process happens W_5.[15] In W_5, the five minutes following Z_1 involve the following process: the cooked egg uncooks by transmitting heat to the pan and coalesces into a raw egg; air molecules form waves converging on

[14] The argument can be run with other deterministic theories, but this needlessly complicates things.

[15] Z_1 and Z_0 are complete states of the world at a time, so given bi-directional determinism there is only one such physically possible world.

the center of the pan, as a result of which the egg leaps upward and a shell closes around it. Obviously we never observe processes like this one in our world. The reason is that this is an *anti-thermodynamic* process, and the second law of thermodynamics says that those are extremely unlikely to happen. (More on this in the next section.) But anti-thermodynamic processes are nonetheless physically possible. Now, as Elga notes, one crucial feature of anti-thermodynamic processes is that they are extremely sensitive to initial conditions, in the following sense. Physically possible microstates of the world can be represented as points in phase space, a continuous space with six dimensions for each particle (one for each coordinate of the particle's position and momentum). Trajectories in phase space represent how possible microstates evolve over time and volumes in phase space represent sets of possible microstates. Let COOKED be the set of states that are exactly like Z_1 with respect to coarse-grained macroscopic parameters like temperature and pressure. In all of those states a cooked egg sits in a pan, Gretta has memories of having cracked it, etc. Some of the states in COOKED (e.g. S_1) have thermodynamically normal futures in which the egg remains in the pan, gradually cooling down. Others (e.g. Z_1) have thermodynamically abnormal futures: they evolve into states in which five minutes later the egg is in its shell, uncooked. Let's call AB the set of such abnormal states. One can give plausible statistical-mechanical arguments for the following two claims. First, AB occupies only a tiny portion of the volume occupied by COOKED. This means that almost all the states near Z_1 in phase space have thermodynamically normal futures. Second, the sets in AB are scattered throughout the volume occupied by COOKED. This means that even the slightest change to the positions or velocities of a few of the molecules making up the pan or the egg in Z_1 would yield a state in which the egg remains cooked 5 minutes later. This is exactly the kind of change that makes for a small miracle. Now consider a world (call it W_6) that has the following characteristics. W_6 evolves exactly like W_5 until Z_1. At this time, a small miracle happens that changes the positions of a small bunch of the molecules in the pan, after which W_6 evolves lawfully again. Given the extreme sensitivity of the Z_1-to-Z_0 process to initial conditions, this means that in W_6 the egg remains in the pan, cooling down. Finally, let W_7 be the temporal reverse of W_6. This

world starts in a very different state from ours until we reach a time at 8:05am; at this time a small miracle happens that shifts a few of the molecules around, after which W_7 evolves exactly like our world. This small miracle is a *convergence* miracle: it yields a world in which Gretta doesn't crack the egg but history is exactly similar to ours at 8:05 and after.

Now that we have built W_7, we can answer the question above of how traces of an event can lawfully appear in the absence of the event. Consider the state of W_7 at 8:05am and run it backward. At 8:05am W_7 matches the actual world except for a small region in which the miracle occurs. As we run time backward, this region of discrepancy rapidly expands. Inside that region, things look statistical-mechanically typical. (For instance, the egg rots as time gets earlier.) Outside the region, events look statistical-mechanically atypical. (For instance, eggs get less rotten as time gets earlier.) At 8:00am, the infected region includes the egg sitting on the pan. Over the next five minutes, the egg becomes progressively warmer, and by 8:05am it is in a state that suggests it has been recently cracked in the pan and cooked. But this suggestion is entirely misleading: long ago the egg formed as a puddle of rotten slime, and reached its cooked state by a process of reverse-rotting. This argument generalizes show that other misleading traces of Gretta having cracked the egg arise through similar anti-thermodynamic processes.

W_7 also shows that there is no asymmetry of miracles. Despite the fact that in the actual world there are many traces of Gretta having cracked the egg, it only takes a small miracle to get a world in which Gretta doesn't crack the egg but all the traces of her having done so are there. It doesn't take a bigger miracle to get a converging world than a diverging one. This has disastrous consequences for Lewis, since it entails that his account cannot capture the counterfactuals that matter for strategies and decision-making. Gretta cracking the egg is an effective strategy for cooking the egg. But Lewis's account doesn't give us this result. Since W_7 is a very close world by Lewis's standards (it contains only a small miracle that buys an enormous region of perfect match), the counterfactual 'If Gretta hadn't cracked the egg the egg wouldn't have been cooked five minutes later' comes out false.

2 The Statistical-Mechanical Account of Causal Dependence

Lewis's account of difference-making faces severe difficulties. Recently, however, a more promising account along Lewisian lines has been defended by Albert (2000, 2013), Kutach (2002, 2007) and Loewer (2007, 2012).[16] Albert, Kutach and Loewer ('AKL' for short) retain the guiding idea behind Lewis's theory, viz. that c makes a difference to e when two situations that are as similar as possible to the actual world but differ with respect to c's occurrence also differ with respect to e's occurrence. AKL also retain the Lewisian idea that the asymmetry of difference-making is a product of an asymmetry of traces. However, they evade the Albert-Elga objection by making causal dependence explicitly sensitive to statistical-mechanical facts. For this reason I will follow Loewer (2007) in calling their proposal the *statistical-mechanical account* of causal dependence.[17]

2.1 Statistical-Mechanical Conditional Probabilities

One of the main differences between AKL and Lewis is that AKL's theory appeals to objective probability in both the deterministic and indeterministic case.[18] More precisely, AKL endorse two significant theses. The first is that a proper account of the causal dependence of e on c must appeal to the *objective conditional probabilities* of e given c (and perhaps given other events as well). The second is that the relevant probabilities come from statistical mechanics. These are two theses that my own account of causal dependence will endorse as well. In fact, I think there are good reasons to think that *any* account of causal dependence must endorse them. My goal in this subsection is to make this claim plausible by summarizing AKL's arguments for these two theses.

To start, let's come back to the Albert-Elga objection. Remember that in the world

[16] For present purposes it will be harmless to ignore the various differences between these authors.
[17] By contrast to Lewis, AKL do not explicitly attempt to provide truth-conditions for ordinary language counterfactuals.
[18] Lewis's account appeals to objective probabilities in the indeterministic case only.

that creates troubles for Lewis's account, W_7, the misleading traces of Gretta having cracked the egg are formed through anti-thermodynamic processes. So a natural suggestion to solve the problem is to make the notion of similarity relevant for counterfactual evaluation sensitive to the facts that account for the absence of such processes in our world, in order to make W_7 count as more distant from the actual world than thermodynamically normal worlds.

To flesh out this suggestion we need to look at the standard statistical-mechanical explanation of the absence of anti-thermodynamic processes in our world. This explanation has its roots in the work of Boltzmann. Consider an instance of thermodynamically normal behavior, say the fact that an ice cube floating in a cup of warm water will melt in the next minutes. And take the volume of phase space made up of all the microstates compatible with the ice cube floating in the water. Boltzmann was the first to show the following fact, already mentioned in the previous section: on the standard Lebesgue measure over phase space, the set of microstates that sit on thermodynamically abnormal trajectories corresponding to the ice cube *not* melting in the next few minutes is incredibly *tiny*. This is the first step of his explanation. His second groundbreaking step was to construe the Lebesgue measure as a *probability measure* specifying the probability that a system is in a microstate **m** given that it is in a macrostate **M**. (This implies e.g. that if two regions of phase space compatible with the macrostate of a system are equal in size on the standard measure, the system is as likely to be in one as in the other.) These two steps together imply that it is enormously *unlikely* that the ice cube won't melt in the next few minutes. More generally, what follows from these two steps is that thermodynamically abnormal behavior is incredibly unlikely to happen: no wonder, then, that we never observe such behavior. However, there is a crucial problem with this explanation known as the *reversibility objection*. Given the time-symmetry of fundamental physical laws, the explanation works equally well in reverse. Consider an ice cube half-melting in a cup of water over the course of five minutes, until time t. Boltzmann's argument entails that at time t, the ice cube is enormously likely to have spontaneously formed in the water during the last few minutes; in other words, it entails that the actual evolution of the ice cube is

incredibly unlikely. Thus Boltzmann's argument cannot explain the *time-asymmetry* of thermodynamic behavior: it doesn't explain why anti-thermodynamic processes are incredibly rare from past to future but extremely frequent in the reverse temporal direction. To explain this, we need to supplement Boltzmann's statistical argument with an asymmetric boundary condition: we need to posit that the universe actually began in a macrostate of very low entropy. Borrowing a phrase from Penrose, Albert (2000) calls this the 'Past Hypothesis' (PH). On this view, the true statistical-mechanical probability distribution is not the Boltzmannian uniform distribution over phase space. Rather, it is the uniform distribution *conditional on the PH*. Let's call this probability distribution *SMP* (for 'Statistical-Mechanical Probability'). The main reason for taking *SMP* as the correct statistical-mechanical probability distribution is that it blocks the reversibility objection. On *SMP*, the probability that the entropy of an isolated system was lower earlier is close to 1, since holding fixed the very low entropy of the initial state of the universe almost all the microstates compatible with the current macrostate of the system sit on trajectories corresponding to the system having lower entropy in the past (Albert, 2000). *SMP* does more than inducing a probability distribution over thermodynamic propositions (e.g. propositions about the temperature of a small region). It also induces a probability distribution over all propositions that supervene on microphysical histories. For instance it assigns a probability to the proposition that the outcome of a specific fair coin toss will land heads; it does so by measuring the proportion of microphysical trajectories that instantiate the coin landing heads among the set of microphysical trajectories that instantiate the coin being tossed. Likewise, *SMP* plausibly assigns probabilities to various chemical, biological, meteorological etc. phenomena. More generally, on the assumption that macro-events supervene on microphysical history, *SMP* induces a probability over all macro-events, thereby providing a kind of probabilistic map of the world.

These considerations suggest that perhaps Lewis's account can be insulated from the Albert-Elga objection if we amend the similarity metric and add conformity to PH as a respect of similarity. But as Loewer (2007, 315-6) notes, this will not work. There are worlds very much like W_7 which obey PH. Such worlds start in a state of very

low entropy and evolve in a thermodynamically normal way for a while. But at some point before 8:05, a statistical fluke happens and traces of Gretta having cracked the egg start appearing through thermodynamically abnormal processes, without Gretta having cracked the egg earlier. Shortly before 8:05 a small miracle happens so that after 8:05 those worlds are exactly similar to ours. *SMP* assigns a very low probability to such worlds, but they are nonetheless physically possible.

There is a deep issue with Lewis's approach to counterfactuals lying in the offing there. Statistical-mechanical considerations suggest that for any macroscopic event c, there are always a few microstates instantiating c that evolve into extremely abnormal situations. For instance, it is physically possible for a rock thrown at a window to suddenly evaporate into thin air, or for an egg to jump spontaneously out of a pan. In light of it, it is plausible to think that for many intuitively true counterfactuals with a non-actual macroscopic antecedent, there is always a physically possible world in which the consequent is false because of such a bizarre evolution. (And as Loewer points out such a world may well obey PH.) For instance, there is a physically possible world in which Suzy throws the rock but the rock evaporates into thin air and the window remains intact. Moreover, the sorts of considerations marshaled by Albert and Elga show that in at least many of those cases, the bizarre world in question will be no less similar than a world in which both the antecedent and the consequent is true. This casts doubt on the whole idea of evaluating counterfactuals in terms of similarity between worlds. We may call this *the problem of macroscopic events*.

The problem of macroscopic events isn't a problem for Lewis only. Any plausible solution to Russell's problem must permit macroscopic events to be causes to explain effective strategies. (We can choose to perform a certain macroscopic event to promote a desired outcome, but we cannot choose to make a precise microscopic instantiation of the macro-event occur.) And as already noted any such solution must connect causal dependence to physical laws. But given a macroscopic event c, the laws of nature allow many different possible evolutions, including very bizarre ones. The most one can say given a macroscopic description M of the world at a time is that given the statistical-mechanical probability distribution over the microstates compatible with M certain

evolutions are much more *statistical-mechanically probable* than others. This suggests that any plausible account of the causal dependence of e on a macroscopic event c will have to appeal to facts about the statistical-mechanical probability of e given c (and perhaps other events as well), even if determinism is true. We need conditional probabilities to relate macroscopic descriptions of the world to the sort of microscopic descriptions that laws of nature take as inputs.[19]

2.2 The Nature of Statistical-Mechanical Chances

Before we look at AKL's account it will be useful to say a few words about the nature of statistical-mechanical conditional probabilities (SM-probabilities for short). If as I argued any account of causal dependence will have to appeal to them, we need to get clear on what they are exactly. I'll start with two pressing issues raised by the appeal to SM-probabilities. First, we need to make sure that those probabilities are not metaphysically dependent on causal facts. Otherwise an account that appeals to those probabilities won't be reductive and therefore won't work as a physicalist solution to Russell's challenge. The second issue is that SM-probabilities are often taken to be subjective probabilities (degrees of ignorance). If correct, this threatens to make our account of difference-making unacceptably subjective.

Let's consider the second issue first. There is a very strong reason to take SM-probabilities to be *objective* probabilities rather than mere degrees of belief. Those quantities play an indispensable role in explaining the second law of thermodynamics. But it is hard to see how degrees of ignorance could play such an explanatory role (Albert, 2000). What could our ignorance have to do with whether a physical quantity like entropy decreases or increases over time? The main reason for taking SM-probabilities to be subjective is that non-trivial SM-probabilities are compatible with determinism[20], whereas non-trivial objective probabilities are often said to be incompatible with determinism. Lewis famously endorses this view: 'To the question of how chance can be reconciled with determinism, ... my answer is: *it can't be done*' (1986a, 118). But

[19] See Ismael (2009) for similar considerations.

[20] Indeed the standard Boltzmannian explanation of statistical-mechanics presupposes classical mechanics.

Lewis's incompatibilism is predicated the assumption that all *bona fide* objective probabilities are dynamical probabilities - probabilities encoded in the dynamical laws of our world. If this assumption is correct, then incompatibilism is plausible. If the dynamics of coin tosses is deterministic, then the chance of the coin landing heads must really be 0 or 1, not $\frac{1}{2}$. However, the assumption that all objective probabilities are dynamical is unmotivated. One may also ascribe objective probabilities to initial conditions, and such chances are compatible with a deterministic dynamics. Indeed, as we will see on Lewis's account of chance it is easy to make sense of such initial probabilities. So there are good reasons to take SM-probabilities to be objective. Indeed, since they attach to singular events they deserve to be called objective *chances*.

Let's now turn to the question of the metaphysical status of those chances. There the issue is that about those chances better not be dependent on causal facts. Here we need to look at the main options in the metaphysics of chance. There are two main theories of objective chance, propensity theory and the best-system account.[21] Propensity theory regards chance as a disposition or tendency of a physical situation to produce a certain outcome. For instance, the proposition that a fair coin has a $\frac{1}{2}$ chance of landing heads means that tossing a fair coin has a certain measurable to land heads. As this gloss of the theory makes clear, propensity theory appeals to causal or causal-like vocabulary to explain chance, so if this theory is true chance is not the sort of thing that can figure in a reductive account of causal dependence. This means we must proceed on the assumption that propensity theory isn't the right metaphysics of chance (at least not of SM-chances). This isn't a very risky move as propensity theory faces severe problems anyway (Eagle, 2004). For instance, propensity theorists have trouble explaining why the relevant tendencies can be measured by probability functions.

The best-system account of chance, which is the one that AKL adopt, was developed by Lewis (1994). Lewis based his theory of chance on his Humean theory of laws of nature, according to which laws of nature are the theorems of the best systematization

[21] Lists of theories of objective chances usually include actual and hypothetical frequentism as well. But these theories suffer from crippling problems (Hájek, 1997, 2009), so much so that they cannot really be taken seriously anymore. The best-system account, which as we will see retains some key elements of frequentism, fares much better.

of the fundamental physical truths of our world. In turn, the best systematization is the true theory that best combines the theoretical virtues of simplicity and strength or informativeness. These two virtues compete: 'an uninformative system can be very simple; an unsystematized compendium of miscellaneous information can be very informative (1994, 474). The best theory is the one that balances these two virtues optimally. (There is of course no guarantee that our world has a unique system that is good enough to deserve the title of best system, but Lewis argues that the successes of physics may let us hope that there is.) So far chance isn't in the picture. But Lewis extends his account of laws by letting theories that use probability functions enter the competition for the status of best system. For now we cannot speak of those theories as being true since we haven't yet said what the probability functions represent. Instead we introduce a third theoretical virtue, probabilistic informativeness or *fit*: the higher the likelihood of the actual pattern of particular physical facts according to the theory, the higher the theory's fit. Now we let theories compete according to how well they balance simplicity, strength and fit. The theorems of the best theory are the laws, some of which are probabilistic. The chances are the probabilities determined by these probabilistic laws. The underlying idea is that by specifying probabilities a theory may become much more informative while remaining relatively simple. For instance, a theory that describes the actual pattern of heads and tails (HTHTTHHTTHHHT...) among tossed coins would be very informative but not very simple. A theory that includes a probability of $\frac{1}{2}$ for tossed coins landing heads is much simpler and still tells us something informative about the actual frequencies of heads and tails. Since chances are determined entirely by the actual distribution of fundamental physical properties across spacetime, Humean chances are not dependent on causal facts.

The best-system approach bears many similarities to actual frequentism, the view that the chance of e given c simply is the actual frequency of es among cs. Since chances earn their name by having a high degree of fit with the actual distribution of fundamental physical properties, it is a built-in constraint that chances cannot deviate too much from actual frequencies. But the best-system approach escapes many of the problems that cripple actual frequentism. For instance, actual frequentism entails that

chances are never irrational, since relative frequencies themselves never are. But one finds irrational chances in quantum mechanics. The best system approach escapes the problem since for reasons of simplicity and coherence with other laws the best system of our world may well contain a law that (e.g.) assigns a probability of $\frac{1}{\sqrt{2}}$ to an electron having spin up, even if this doesn't exactly match the actual frequencies. Another problem is that frequentism implausibly entails that a fair coin that is never actually tossed has no well-defined chance of landing heads if tossed. On the best-system approach, by contrast, considerations of simplicity, symmetry and so on can make it the case that chances for events that never or seldom occur are nonetheless well-defined. Moreover the best-system approach doesn't face the difficulties that plague propensity accounts. For instance, the best-system approach has no problem explaining why chances are can be measured probability functions, since it is built into the account that chance is (a distinguished kind of) probability function. Overall, there are good reasons to regard the account as a very promising metaphysics of chance. This isn't to say that the best-system approach is without problems. Simplicity is relative to a language, so there is an issue of selecting the right language to formulate candidates to the status of best system. And as Elga (2004) notes, Lewis's notion of fit is problematic in infinite universes. For instance, if the actual sequence of coin tosses is infinite, it seems that any theory will assign a probability of 0 to this sequence. How best to formulate the best-system approach so as to escape this objection is very much a matter of ongoing debate, one that I won't try to settle here.[22] In what follows I will simply assume that some version of the best-system account of chances is true.

It is important to separate the best-system account from two further assumptions about chance endorsed by Lewis. These assumptions are not entailed by the best-system account of chances (although they are compatible with it). The first is the assumption that all chances are *dynamical* chances. As Loewer (2004) points out, if one adopts the best-system account of chances there are no reasons to limit chances to dynamical chances: one may also accept the existence of chances that attach to initial conditions. Adding to the deterministic dynamical laws a probability distribution over

[22] See Loewer (2004) and Hoefer (2007) for promising proposals.

initial conditions may make the theory substantially more informative while costing little in strength. Thus if one adds a probability distribution over initial conditions to Newton's laws (and the PH), one can capture statistical-mechanical regularities of our world. Likewise, if one adds the so-called quantum equilibrium distribution to the dynamical laws of Bohmian mechanics, one can recapture the statistical quantum regularities described by standard quantum theories. Indeed, it is in fact one of the main advantages of the best-system approach over its competitors that it can easily make sense of such probability distributions over initial conditions. And since initial condition chances are entirely compatible with determinism the best system approach also has the advantage of explaining how statistical-mechanical and other objective chances can exist in a world whose dynamics is deterministic. The second extra-commitment that is relevant here is Lewis's assumption that it is in the nature of chance that the past is never chancy. That is, an event never gives anything other than a 0 or 1 chance to possible events in its past. To represent this, Lewis indexed the chance function to a time, and proposed that for any event c earlier than t, the chance of c at t is either 0 or 1. It is obvious that one should not accept this assumption if one wants to use chance to explain difference-making. For one of the goals here is to explain the time-asymmetry of causation. To do so, one shouldn't oneself to such a temporal asymmetry of chance, since this asymmetry is itself mysterious in light of the symmetric character of the laws of physics. (This isn't to deny that in our world there are important differences between past and future-directed chances, but such asymmetries need to be explained from the physics, not presupposed from the start.) Once we drop the assumption that the past isn't chancy, there is no reason to index chances to times.

While we are on the topic of indexes for chance, let me note that plausibly chances should be indexed to *reference classes* - a move that neither Lewis nor AKL make. Here is why. Consider $SMP(w/s)$, the SM-chance of the window breaking given that Suzy throws the rock. How can we acquire evidence about its value? After all, we are not really in a position to inspect the parts of phase space instantiating s and counting the proportion of them that lead to w. Rather, the natural answer is that we can learn about it by observing associated frequencies, for instance the frequencies of rock-throwings

that are followed by episodes of window-breaking in the vicinity of the throw. But such frequencies are not direct evidence for single-case chances. Rather, they provide direct evidence about *repeatable chances* or *general probabilities* - probabilities that attach to event-types rather than singular events (Ismael, 2011). For instance, the aforementioned frequencies provide direct evidence about the general probability of a window breaking given that a rock is thrown at it. But s and w are each members of many event-types; for instance, s is an instance of a rock being thrown, of a rock being thrown by Suzy, of a rock being thrown with such-and-such velocity, and so on. Likewise, w is an instance of a window breaking, a window breaking in such and such a way, and so on. Which of the corresponding frequencies is relevant for assessing $SMP(w/s)$? This is a version of the famous *reference class problem*. An attractive way to get out of the problem is to follow Hájek (2007) and explicitly relativize chances to reference classes for the relevant singular events. On this view, there is no such thing as $SMP(w/s)$ *simpliciter*. Rather, $SMP(w/s)$ is relative to the specification of reference classes W and S for w and s respectively, and is equal to the general probability of W given S. We may denote this relativized chance with '$SMP_{W,S}(w/s)$'. This solves the problem because it is now obvious how we can acquire evidence about chances from frequencies. Frequencies give us evidence about the values of certain general probabilities, which are in turn equal to the values of chances relative to the relevant reference class. This relativization will be assumed in what follows, although in most cases when I speak of a particular chance I will leave the reference classes implicit, as it should be obvious from the context.

2.3 AKL's Solution to Russell's Problem

Let's now come back to AKL's solution to Russell's problem. In a nutshell, AKL's idea is that a cause c must raise the SM-probability of its effect e's occurrence, conditional on the macrostate of the world outside of the region of c's occurrence.[23] More precisely:

> **AKL**. c makes a difference to e just in case $SMP(e/c.m_w)$ is substantially higher than $SMP(e/{\sim}c.m_w)$.

[23] In some versions (e.g. Loewer (2012)) the relevant state is the *micro*state of the world outside of the region of c's occurrence. As far as I can see this makes no substantial difference.

where m_w is the macrostate of the world outside of the region of c's occurrence, at the time of c's occurrence.

AKL retains the Lewisian idea that to evaluate whether c causes e one should compare situations that differ with respect to c's occurrence but are otherwise as similar to actuality as possible. But here the relevant respects of similarity are not determined by Lewis's metric. Instead, the worlds we should consider are worlds that (a) obey the same laws as ours, (b) start in a macrostate of very low entropy and (c) are as much macroscopically similar to our world as possible at the time of c's occurrence. This gives us a solution to the localization problem very similar to Lewis's. Given the state of the actual world (in particular the absence of backup), Suzy throwing the rock raises the probability of the window breaking. This is entirely compatible with the existence of nomologically possible situations in which Suzy throws the rock but the window doesn't break. Note that by contrast to Lewis's, AKL's respects of similarity are not intended to single out a unique c-world and a unique $\sim c$-world. Rather, they select two sets of worlds: difference-making is assessed by checking whether the proportion of e-worlds is higher among the relevant c-worlds than among the relevant $\sim c$-worlds on the standard Lebesgue measure.

AKL's explanation of the direction and time-asymmetry of causal dependence is similar to Lewis's but escapes the Albert-Elga objection. On their view just as on Lewis's, the direction and time-asymmetry of causal dependence is a product of a contingent physical asymmetry of traces: the fact that events tend to leave traces in their future but not in their past. But AKL do not construe this asymmetry as an asymmetry of overdetermination. Instead, the asymmetry of traces is a statistical-mechanical asymmetry. Traces of e do not nomologically determine e; rather, what makes them traces is that they make e very likely on the statistical-mechanical probability distribution. For instance, traces of Gretta having cracked the egg (her memories, the broken shell, and so on) do not nomologically determine Gretta having cracked the egg since they could have arisen, via an anti-thermodynamic process, without Gretta having cracked the egg. But since on *SMP* such processes are incredibly unlikely, the probability of Gretta having cracked the egg given the presence of those traces is very high. When we

plug the asymmetry of traces into **AKL**, we get the result that effects do not make a difference to their causes, and that causes normally come before their effects. When e is in the past of c, the probability of e conditional on m_w must remain very high whether or not c occurs. The reason is that m_w contains many events conditional on which e has a very high chance of having occurred, so that little room is left for c to make a difference to e's probability. By contrast when e is in c's future, m_w will typically not contain any traces of e, which makes it possible for c to make a large difference to e's probability of occurring. For instance, it will turn out that Brian tossing his cigarette in the forest makes a difference to the forest burning later, since the state of the world at the time of the toss doesn't contain traces of the forest burning a few minutes later save for the toss itself. By contrast the forest burning doesn't make a difference to Brian having tossed his cigarette earlier, since the state of the world at the time of the fire will contain many traces of the toss (the cigarette lying on the ground, Brian's memories, and so on). Like Lewis's theory of the causal asymmetry, AKL's explanation leaves a loophole for backward causation in cases where there are no records of a past event.[24] Another advantage of the account is that it provides a beautiful unification of the causal asymmetry and the thermodynamic asymmetry, as both are shown to derive from certain features of the statistical-mechanical probability distribution of our world (which themselves derive from certain features of the initial condition, in particular its low-entropy state).

2.4 Problems for AKL

AKL's account fares better than Lewis's insofar as it escapes the Albert-Elga objection (and the problem of macroscopic events that lies at its source). Nevertheless, it faces its own problems. I will mention three of them.

One worry is that the asymmetry of traces may not be pervasive enough to make the causal asymmetry as strict as we think it is. As Frisch (2007, 2010) points out, the example of Nixon pushing the button is a very special case in that Nixon's action has exceptionally consequences for the fate of the Earth. Focusing on this example may lead

[24] As we will see in the next section one worry is that this leaves too much of a loophole.

us to exaggerate the pervasiveness and persistence of traces. It is easy to find more mundane cases in which traces of a past event are sparse or disappear very quickly. Frisch (2010) takes as a case in point a situation discussed by Albert (2000) of a bunch of ice cubes dropping into glasses of water after sliding down along a Galton board.[25] Whatever records are produced by the ice cubes sliding down the board will disappear pretty quickly. The drops of water along the ice cubes' paths evaporate, I will soon forget the details of the experiment, etc. so that by the next day all the macro-traces of the experiment will have disappeared.[26]

A second problem pointed out by Frisch (2010) is that there are realistic cases in which an event e is a record of a cause c but e's record is under my direct control, so that on AKL's view doing e is an effective strategy to influence the prior occurrence of c. Imagine that while playing a piano piece, I am unsure whether I am playing a part that is repeated in the score for the first or the second time. I know from experience that when I play the piece, my decision to play the second ending is good evidence for my having already played the part once. It seems that in that case whatever decision I make about playing the second part or not will constitute a record of whether I have already played the first ending. And we can imagine that there are no other records in the present macrostate of the world, or that such records are not very reliable (perhaps I have no more than a vague memory of having already played the part once). It seems that in those circumstances, **AKL** entails that my decision to play the second part makes a difference to whether I played the first, so that by deciding to do so I thereby influence what I played earlier. But this is wrong: in this case intuitively my decision to play the second part is evidentially but not causally relevant to what I played earlier. This is a case where **AKL** turns out to be extensionally inadequate.

The third problem is as follows. As we saw, one problem with Lewis's account of difference-making is that it leaves it mysterious why difference-making as Lewis

[25] Albert uses this as an example of a system that starts in a low entropy state - the ice cubes are all collected at the top - that can evolve indeterministically into different macrostates corresponding to different configurations of ice cubes in water glasses.

[26] As Frisch points out, the processes by which these traces disappear are precisely thermodynamic processes. Although AKL may be right that the existence of traces is closely connected to the facts that account for thermodynamic phenomena, the thermodynamic arrow also acts as a destroyer of traces.

construes it should be the relation that matters for rational decision-making. AKL fare better than Lewis in that they attempt to provide a principled explanation for why the relation picked out by **AKL** is practically relevant. But as we will now see their story faces several objections.

AKL attempt to justify the practical relevance of their notion of difference-making as follows (see Albert (2000, 128), Loewer (2007, 316-7) and Loewer (2012, 127)). Start with the assumption that at any time t there are some small parts of the state of the world at t over which I have *direct, unmediated control*. Specifically, I have direct control over a certain range of *decisions* or *actions* at t: an 'ability to freely choose one among alternative decisions independently of anything else in the universe' (Loewer, 2012, 127). Second, it is assumed that the things over which I have direct control in turn allow me to *influence* other events: specifically, a decision occurring at time t allows me to influence an event e to the extent that the decision is correlated with e, *holding certain background facts fixed*. Direct control over a decision can thus be parlayed into influence over other events to the extent that the decision is correlated with the event holding background facts fixed. Third, AKL assume that the relevant background facts are the time-t state of the world, outside of the region of occurrence of my decision. Influence, then, is a matter of correlation between a decision and another event, holding fixed the outside state of the world at the time of the decision. It follows that my decision can influence e exactly to the extent that the decision *makes a difference* (in the sense of **AKL**) to the occurrence of e.

It seems to me that there are two main problems with this story. The first one has to do with AKL's construal of the notion of direct control. Their notion of control is explicitly a libertarian one: we have control over a decision to the extent that we have an ability to make that decision in a way that is not determined by the rest of the universe. As AKL explicitly recognize if determinism is true we do not have direct control over anything in this sense: the supposition that we have control is a *myth* (Loewer, 2012, 127). One worry here is that it isn't clear why we developed this myth in the first place, and consequently why we take ourselves to have influence over anything. If we do not really have control over anything, what is the point or rationale of acting as if we do?

The second problem concerns AKL's notion of influence. Here it is useful to distinguish two ingredients to their conception of influence. First, there is the idea that influence is a matter of correlation: a decision influences an event to the extent that the occurrence of the event is more probable given the decision. Second, there is the thought that the correlation must hold given certain background facts, specifically *given the present state of the world* outside of the region of the decision's occurrence. The first idea strikes me as very plausible. There is something obvious or self-evident about the idea that we can usefully influence the occurrence of a desired outcome to the extent that we can act in a way that is *correlated* with the outcome's occurrence. As Papineau puts it in a somewhat different context: 'Doesn't everybody want it to be probable that they will get what they want?' (2001a, 244). But the idea that the correlations that can be exploited for goal advancement are those that hold *given the present state of the world* is much more problematic, for two reasons.

To get to the first one, let's stipulate that a present decision *influences** an event e to the extent that the decision is correlated with e, conditional on the state of the world *an hour ago*. Presumably we can sometimes influence* events in our recent past. For instance, given that the state of the world an hour ago doesn't contain records of my having had dinner in the last hour, my decision to have dinner right now correlates with and thereby influences* my not having had dinner during the last hour. Now, it makes sense to ask the following question: why is AKL-influence rather than influence* the sort of relation that matters for goal advancement? To answer this question, AKL would have to give a principled justification for the claim that influence should be assessed relative to the present state of the world rather than the state of the world an hour ago. But it is hard to see what such a justification would be. The only one I can find in AKL's work is as follows. At the start of chapter 6 of *Time and Chance* (2000) - the chapter where he discusses influence - Albert makes the simplifying assumption that we have *direct knowledge* of the present state of the world. If that were true, then it would make sense for us to holding the present state of the world fixed when assessing what decisions we should make. This simply follows from the self-evident principle that in decision-making we should assess the likely consequences of our actions

conditional on what we know to be the case. But of course as Albert himself recognizes the assumption that we have direct knowledge of the present state of the world is an extremely unrealistic idealization. At most we only have epistemic access to a very small part of the present state of the world.[27] (Note that this is an instance of the problem of close-by alternatives: why should AKL-influence rather than some close-by alternative such as *influence** be the relation that matters in the context of rational decision-making?)

Here is a second reason why AKL's conception of influence as correlation given the present state of the world is problematic. As we have seen when reviewing the Humean theory of the causal direction, a proper theory of the time-asymmetry of causation shouldn't make backward causation impossible in our world by *fiat*. AKL's theory respects this requirement: the impossibility of backward causation isn't built in AKL's notion of difference-making, but *explained* in terms of the asymmetry of traces. But their conception of influence nonetheless makes *simultaneous* causation impossible by fiat. If what a decision can influence is what it correlates with *given* the present state of the world, then the decision cannot influence any part of the present state of the world since that state is held fixed. But a proper theory of causation shouldn't make simultaneous causation impossible, for precisely the same reasons that it shouldn't make backward causation impossible. As we have seen, one problem with stipulating that backward causation is impossible in our world is that time travel seems physically possible and perhaps even actual. So a theory that stipulates the impossibility of backward causation may not be extensionally adequate, or at least not general enough. (Even if time-travel actually doesn't happen in our world, we would like a proper theory of causation not to depend crucially on this contingent fact.) But by opening the possibility of backward causation time-travel also opens the possibility of *simultaneous* causation. Suppose that at time t I take the decision to press the button on a time

[27] On certain theories of time, there is something metaphysically privileged about the present state of the world: in particular, on the 'moving now' conception of time, the present state of the world is all that exists. If this is right perhaps this could be parlayed into an explanation for why we should hold fixed the present state of the world (rather than some other state) for the purpose of evaluating which goals we can advance. But the idea of the 'moving now' is famously hard to square with physics, in particular with the relativity of simultaneity entailed by special relativity. Moreover this conception of time is very alien to AKL's staunchly four-dimensionalist framework.

machine at t+1; at t+1 the machine sends a particle back in time to t. In this case we would like to say that my time-t decision influences the present state of the world. We shouldn't rule out the possibility of such cases at the outset, but AKL does.

3 Probabilistic Theories of Causal Dependence

Let me briefly summarize the main results of the last two sections. We have looked at two variants of the same idea: that to evaluate whether e causally depends on c, we should compare two situations (or sets of situations) that are similar to each other and actuality in certain respects, but which differ with respect to whether c occurs. The two theories we have seen differ mainly on what they take the relevant respects of similarity to be. For Lewis the relevant respects of similarity are those encoded in [**S1**]-[**S4**]; for AKL similarity is similarity with respect to the present state of the world. And we have seen that both theories face versions of the following two problems. First, they do not seem to give us an entirely satisfactory account of the causal direction and time-asymmetry. Second, both have difficulties explaining why the respects of similarity that matter for difference-making are the ones that should be held fixed in rational decision-making.

Nevertheless, there are important positive lessons to draw from AKL's approach. In particular, we have uncovered strong reasons for thinking that a plausible account of the causal dependence of e on c will have to appeal to the fact that the conditional probability of e is higher given c than given $\sim c$ (perhaps holding certain other facts fixed). One argument for this pertains to what I called the problem of macroscopic events. Given that the throwing of a rock can lead to bizarre evolutions in which (e.g.) the rock suddenly dissolves into thin air and the window doesn't break - even if we hold fixed the state of the world outside of the region of rock-throwing, the most one can say is that the window is more likely to break given that the rock is thrown. So plausibly any account of causal dependence will have to appeal to such conditional probabilities. Moreover, as we have seen when reviewing AKL's theory of influence, the idea that influence is a matter of correlation between an action and desired outcome has an air

of self-evidence to it. So the idea that an effect is more likely given its cause than given its non-occurrence seems to be the right starting point for a theory of effective strategies. Finally, if correlation rather than determination is the mark of causation, there is no problem of localization. The existence of correlations between localized events is perfectly compatible with the globality of physical laws. For instance, the fact that Suzy throwing the rock and the window breaking are correlated is compatible with the fact that the throw by itself doesn't determine the window to break.

Our next order of business is therefore to look at *probabilistic theories* of causal dependence, which take as their starting point the idea that a cause c is correlated with its effect e - i.e. that $P(e/c) > P(e)$. Such views face two challenges. The first is that correlations can hold between events that do not stand in causal relations. For instance, the reading on a barometer is correlated with the occurrence of a storm later. A storm is more likely to occur later if the reading is low than if it is high. The reason isn't that barometer readings cause storms; rather, the correlation arises because both events have a common cause, namely low atmospheric pressure. So one challenge for probabilistic theories is to distinguish genuine causal correlations from *spurious* correlations. Moreover the idea that cause and effect are correlated doesn't give us an asymmetry between them. Correlation is symmetric: it is a theorem of the probability calculus that $P(b/a) > P(b)$ just in case $P(a/b) > P(a)$.[28] So a second challenge for probabilistic theories is to explain where the causal direction comes from

[28]To see this, note that we have $P(b/a) > P(b/{\sim}a)$ just in case

$$\frac{P(b\&a)}{P(a)} > \frac{P(b\&{\sim}a)}{P({\sim}a)} \tag{2.1}$$

From (2.1), it follows that

$$P(b\&a) > P(b)P(a) \tag{2.2}$$

and thus that

$$\frac{P(b\&a)}{P(b)} > P(a) \tag{2.3}$$

But this is just to say that

$$P(a/b) > P(a) \tag{2.4}$$

(and why it is typically aligned with the direction of time).

The probabilistic theories I will discuss all rely on the same idea to solve these two challenges. The idea there is that cause and effect are distinguished not by their probabilistic relations to one another (since those are symmetric) but by their probabilistic relations to other events. A bit more precisely, the idea is that events in our world are related by probabilistic relations which together form a *probabilistic network*, and that this network has certain features (in particular certain *asymmetric* features) on the basis of which one can distinguish causal from spurious correlations and explain the asymmetry between a cause and its effect. We will look at two ways in which this idea has been developed. The first one is Reichenbach's theory of causal dependence, the first sustained attempt to develop a probabilistic account of causation. As we will see, although Reichenbach's account suffers from severe defects, it also gives us various tools and ideas that are crucial for any probabilistic theory of causation. The second one is based on influential work in the Reichenbachian tradition due to Spirtes, Glymour and Scheines (2000) and Pearl (2009). Although these authors are concerned with methodological issues, their central tools and ideas have been used to articulate a metaphysical theory of causal dependence. We will look at the most sustained attempt to do so due to Papineau (1993, 2001b).[29]

The authors in the probabilistic tradition explain difference-making and its asymmetry in terms of objective probabilities, but they often say very little about what the relevant objective probabilities are supposed to be. This is unsatisfactory, especially if we regard the theories I will discuss as solutions to Russell's problem. For a proper solution to Russell's problem needs to explain how the material it appeals to (here probabilities) connects with fundamental physics. The natural suggestion here is to take the relevant probabilities to be statistical-mechanical probabilities as interpreted by AKL, and this is indeed what I will do here. Thus in what follows the symbol 'P' should be interpreted as referring to the statistical-mechanical probability function.

[29] Other important authors in the probabilistic tradition are Suppes (1970), Kvart (2001), and Glynn (2009). Those authors, however, are more concerned with actual causation than with causal dependence, and all of them endorse the temporal theory of the causal direction, which as we have seen is unsatisfactory. Spohn (2001) endorses an account of causal dependence in the probabilistic tradition but his account also presupposes that causes must precede their effects.

This is appropriate because the link between SM-probabilities and fundamental physics is itself transparent and (at least on the Humean interpretation of these probabilities) their existence doesn't depend on causal facts. Moreover, on probabilistic theories the relevant probability function must be very extensive. For all or almost all pairs of events there is a fact of the matter as to whether they are causally related. Since on probabilistic theories causation is a matter of probabilities, this means that for almost all pairs of events the probabilities relating these events must be well-defined. *SMP* answers the challenge since as we have seen it gives us a very extensive probabilistic map of the world.

3.1 Reichenbach's Theory of Causal Dependence

Reichenbach (1956) was the first to propose a sophisticated probabilistic theory of causal dependence. His analysis proceeds in two steps. In the first stage, he offers a preliminary theory of causal dependence that appeals to the direction of time to fix the direction of causation and to distinguish causal from spurious correlations. In the second stage, references to the temporal order are eliminated, and the causal asymmetry is explained as a product of a contingent probabilistic asymmetry, the *fork asymmetry*.

Reichenbach's preliminary solution to the problem of the causal direction is a version of the Humean theory of the causal direction, on which what distinguishes causes from effects is simply that the former occur earlier than the latter. In addition to the difficulties already discussed, one problem with this view is that it doesn't distinguish causal from spurious correlations. The barometer reading comes earlier than the storm and is correlated with it, but doesn't cause it. To deal with the problem of spurious correlations, Reichenbach appeals to the fact that common causes *screen off* their independent effects from each other - a fact which he was the first to point out. In general, an event d screens off e from c just in case the probability of e is the same given d and given c and d: $P(e/d) = P(e/c.d)$. That is, d screens off e from c insofar as given d, c and e are probabilistically independent of each other. For instance, the drop in atmospheric pressure screens off the correlation between the barometer reading

and the storm's occurrence. Once the occurrence of the drop is held fixed, the correlation between the two effects disappears. More generally, Reichenbach formulated a principle which he called the *principle of the common cause*. Suppose that two events a and b are correlated, so that $P(b/a) > P(b/\neg a)$, but that neither is a cause of the other. Reichenbach maintained that in this case there is a common cause d satisfying the following conditions:

1. $P(a/d) > P(a/\neg d)$
2. $P(b/d) > P(b/\neg d)$
3. $P(a/b.d) = P(a/d)$
4. $P(a/b.\neg d) = P(a/\neg d)$

That is, the common cause is correlated with both a and b, and it screens off the correlation between them.[30]

The principle of the common cause suggests that we can deal with the problem of spurious correlations by requiring true causal correlations to not be screened-off by a third event. Since the barometer-storm correlation is screened off by the drop in atmospheric pressure, this rightly counts the correlation as a spurious one. The problem with this suggestion is that causal correlations can also be screened off, because - as Reichenbach realized - causal intermediaries screen off correlations between their causes and their effects. For instance, suppose that smoking causes lung cancer solely by affecting the amount of tar in one's lungs. Then holding fixed the amount of tar in her lungs, whether the person was a smoker earlier is independent of whether she will get lung cancer later. To distinguish common causes from causal intermediaries, Reichenbach once again appeals to the direction of time. The idea here is to require that causal correlations not be screened-off by any *temporally prior* third event. Since causal intermediaries are temporally between their causes and their effects, this temporal condition ensures that we won't mistakenly count causal correlations as spurious because they are screened off by causal intermediaries. Thus, Reichenbach's preliminary theory of causal dependence can be formulated as follows:

[30] As we will later see there are a number of circumstances in which the common cause principle fails. I will discuss these circumstances in the next section when I discuss a closely related principle called the 'causal Markov condition'.

Reichenbach (Preliminary). e causally depends on c just in case

1. c happens earlier than e.
2. $P(e/c) > P(e)$.
3. There is no third event d such that
   (a) $P(e/c.d) = P(e/d)$
   (b) d happens earlier than c and e.[31]

Note that on the view of chance I proposed in §2.2, chances are relative to a reference class. To which reference classes, then, should conditions (2) and (3.1) be taken to hold in **Reichenbach (Preliminary)**? The natural hypothesis is that there need only be one reference class for c and one reference class for e relative to which (2) holds, and that for any event d earlier than c and e, there is no reference class D such that 3.1 holds relative to C, E and D. This is the interpretation of such statements that I will adopt in what follows. That is, a statement that some events stand in a certain probabilistic relation should be understood as shorthand for the claim that there are reference classes for these events relative to which the relevant probabilistic relation holds. And the statement that no event satisfies a certain probabilistic condition should be understood as shorthand for the claim that there is no reference class for this event relative to which the relevant probabilistic relation holds.

Reichenbach recognized some of the limitations of the Humean theory of the causal direction. In the second stage of his analysis, he sought to give an analysis of the causal direction that makes it independent of the temporal order. He argued that the direction of causation could be explained in terms of a pervasive statistical asymmetry which he called the *fork asymmetry*. To explain the fork asymmetry we need to introduce the concept of a *conjunctive fork*. Three events a, b and d are said to form a conjunctive fork when they satisfy the equations (2)-(6) above. (That is, each event is correlated with the two others and d screens off the correlation between a and b.) d is the prong of the fork, and a and b are its tips. If d occurs earlier than both a and b and there is no event later than a and b satisfying (2) to (6), d, a and b are said to form a conjunctive fork *open to the future* (see Fig. 2.1). Likewise, if d occurs later than both a and b and

[31] Suppes (1970) offers essentially the same theory of causation.

there is no event earlier than a and b satisfying (2) to (6), we have a conjunctive fork *open to the past* (see Fig. 2.2). If an earlier event d and a later event f both satisfy conditions (2)-(6)), we have a closed fork.

What Reichenbach calls the fork asymmetry is the fact that in our world there are many conjunctive forks open to the future but none or very few open to the past. For instance, the correlation between the barometer reading and the storm is screened off by the earlier drop in atmospheric pressure, while there is presumably no later event that screens off the correlation as well. Thus the drop, the barometer reading and the storm form a conjunctive fork open to the future. And it is not hard to find other actual examples of such forks. Most cases that we would intuitively describe as situations where a common cause induces a correlation between two independent effects will be cases where we have a conjunctive fork open to the future. By contrast it is much more difficult to come up with realistic cases of conjunctive forks open to the future.[32] In general, when two events are correlated with a third one in their future, they are not themselves correlated with each other, or at least the third event doesn't screen off the correlation. For instance[33], smoking and working in an asbestos factory are both correlated with lung cancer later. But just because of this we shouldn't expect the proportion of smokers to be higher among asbestos workers than in the general population. And if both were correlated, we shouldn't expect the correlation to disappear given lung cancer, except via a statistical accident.[34]

Thus, the fork asymmetry is a pervasive asymmetry characterizing the network of probabilistic relations relating events in our world. Note that the fork asymmetry is intimately tied to the thermodynamic asymmetry, and is presumably a product of the same facts about the initial condition of the universe and statistical typicality that explain the thermodynamic arrow. One way to see this is to imagine looking at our

[32] Reichenbach thought that closed forks were relatively common in our world. But as we will see in the next section, the general considerations I mention make it plausible to think that in our world it is very uncommon that the correlation between two events is screened off by an event in their common future, so that both closed forks and forks open to the future are very rare in our world.

[33] I borrow this example from Papineau (1985).

[34] Intuitively: since lung cancer isn't the *source* of the correlation there is no reason to expect the correlation to disappear once we hold lung cancer fixed unless there is a statistical fluke. Here I am using causal language ('source') to justify the claim that there is a fork asymmetry, but bear in mind that the asymmetry is a purely statistical one.

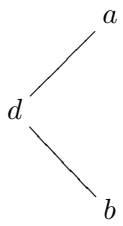

Figure 2.1: Conjunctive Fork Open to the Future

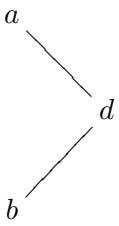

Figure 2.2: Conjunctive Fork Open to the Past

world in reverse. From this perspective we would see lots of forks open to the past, for instance broken fragments of glasses assembling themselves together in a precise alignment to form an unbroken glass (Price, 1996, 119). This is precisely the sort of strange coincidence that never happens in the past-to-future direction, because it is prohibited by the thermodynamic asymmetry and the underlying statistical-mechanical probability distribution.[35] Exactly how to connect the two arrows is not an entirely straightforward matter. But insofar as the connection seems to be there, an account of the causal direction and time-asymmetry that relies on the fork asymmetry is on the right track to eventually provide a unified account of those physical asymmetries.

Reichenbach exploits the fork asymmetry to give an account of the causal direction that doesn't appeal to temporal order. The idea is to replace references to temporal order in **Reichenbach-Preliminary** by references to the temporal direction in which open forks predominate. Thus, rather than requiring causes to come earlier than their effects, his final analysis requires causes to be downstream of their effects with respect to the temporal direction in which open forks tend to be open. Cause and effect are distinguished by their respective places in a global asymmetric probabilistic network. The fact that causation is time-asymmetric is then seen as a consequence of the fork asymmetry - the contingent fact that in our world the predominant direction of open forks is from past to future.

Reichenbach introduced many useful ideas and tools for thinking about the relations

[35] See Horwich (1987, ch. 4) for other suggestive remarks about the connections between the fork asymmetry and the thermodynamic arrow.

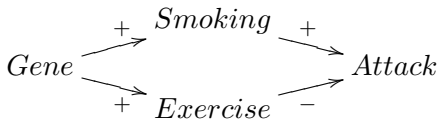

Figure 2.3

between causation and probabilities. But his account suffers from crippling problems. Let me mention three of them. A first problem is that, contrary to what condition (2) of Reichenbach's theory says, mere (positive) correlation is not *necessary* for causation. Cartwright (1979) was the first to make the point with examples of the following sort. Suppose that smoking causes heart attacks, and that in addition there is a gene that causes its bearers to smoke and to exercise a lot. Suppose, moreover, that exercising has a negative causal influence on heart attacks. The causal structure of the case is represented in Figure 2.3. In this example, it may very well be the case that smokers are *less* likely to have a heart attack than non-smokers. If the negative influence of exercising on heart attacks is sufficiently strong, it may well override smoking's positive influence on heart attacks, so that given the correlation between exercising and smoking smokers tend to have heart attacks in lesser proportion than non-smokers. As we might put it, the positive correlation between smoking and heart attacks is *masked* by the correlation between smoking and exercising, and appears only when we hold exercising fixed. (Among people who exercise and among people who don't, those who smoke are more likely to have a heart attack than those who don't.) This case is an instance of *Simpson's paradox*: the fact that a negative correlation between two events can become a positive one if we conditionalize on a third event. It shows that the connections between causation and probability-raising are much more subtle than Reichenbach assumed. Even more, it may appear to spell doom (and has indeed been widely taken to spell doom) for the project of reducing causal structure to probabilistic structure.

The two other problems concern Reichenbach's account of the causal direction and time-asymmetry. Reichenbach thought that instead of explaining the direction of causation in terms of the direction of time (as Hume did), we could explain it in terms of the global direction of open forks instead. But this account runs into the exact

same problems as the ones facing the Humean account of the causal direction. First, Reichenbach's account makes backward causation impossible in our world. Since on Reichenbach's account the temporal direction of any individual causal relation is fixed by the global direction of open forks, his view entails that in our world local instances of backward causation are impossible: any causal process must run in the same direction as the global direction of open forks. Second, Reichenbach's account leaves it mysterious why we can influence the future but not the past. Suppose that my present action c is correlated with a desired outcome e in its past. Why should the fact that c lies upstream of e with respect to the global direction of open forks make it irrational to do c so as to influence the occurrence of e? In other words, Reichenbach's account lacks an explanation of the connection between the asymmetry of effective strategies and the fork asymmetry. It is hard to see what such an explanation could be. Presumably the fact that I cannot use c to influence e depends on much more local factors than the global direction of open forks.

3.2 SGS and Papineau's Theory of Causal Dependence

Although Reichenbach's account suffers from severe problems, there is another probabilistic account of causal dependence very much in Reichenbach's spirit that fares much better (although as we will see it isn't without problems). This account is due to Papineau (1993, 2001b) and Field (2003).[36] Papineau offers by far the most detailed version of it, and I will therefore focus on his work. Papineau's account draws heavily on a range of concepts and techniques inspired by Reichenbach's work and developed by the computer scientist Judea Pearl and his colleagues (see in particular Pearl and Verma (1991); Pearl (2009)) and the philosophers Peter Spirtes, Clark Glymour and Richard Scheines, or 'SGS' for short (Spirtes et al., 2000). The goal of Pearl and SGS is not to articulate a metaphysical theory of causation. Rather, they are concerned with the methodological question of how to infer causal information from purely statistical data. But Papineau argues that their results also establish that causation can be reduced

[36] See also Spohn (2001). (Hausman, 1998, ch. 10) is an excellent critical discussion of Papineau's theory.

to probabilistic structure. For reasons that it would take me too long to go into, the framework of Spirtes et al. (2000) is more congenial to the project of reducing causal structure to probabilistic structure than the framework of Pearl (2009). In what follows I will therefore concentrate on the former work.[37] I will first present the most significant ideas and results of SGS's work. It will be useful here to look at their framework closely, as it contains many ideas that I will use in my own account of causal dependence. I will then turn to Papineau's attempt to use SGS's framework for metaphysical purposes.

3.2.1 Bayes Nets and Causal Graphs

The central device in SGS's framework is a tool (called a *Bayes net*) to represent the dependence and independence relationships encoded in a probability function.

A Bayes net is a kind of *directed acyclic graph* or *DAG* for short. In turn, a DAG is a set of variables and a set of directed edges or 'arrows' between these variables. The values of the variables we'll be interested in represent are propositions representing the occurrence or non-occurrence of singular events. For instance, if the events of interest are John smoking and John getting lung cancer and their contrasts, it is natural to represent them with a variable S that takes value 1 just in case John smokes and 0 just in case it doesn't, and variable L taking value 1 just in case John has lung cancer and 0 if he doesn't. As a convenient shorthand, a probabilistic statement that contains only a variable or a set of variables but no values will be understood as a universally quantified statement over all values of the variables. For instance, we may write $P(Y/X) = P(Y)$ as a shorthand for $\forall x \forall y, P(Y = y/X = x) = P(Y = y)$. We can thus speak of variables being correlated or probabilistically independent of each other The probability distribution for our purposes is the statistical-mechanical probability function. It will also be useful in what follows to speak of a variable X causing another variable Y, which should be read as shorthand for the claim that one of the events corresponding to a value of X causes an event corresponding to the value of Y.

As mentioned, a DAG on a set of variables $\mathbf{V}$ includes directed edges or 'arrows'

[37] Note, however, that many of the theorems in Spirtes et al. (2000) draw on prior work by Pearl and collaborators.

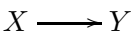

Figure 2.4

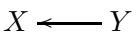

Figure 2.5

Figure 2.6

between the variables in **V**. When the DAG includes an arrow from X to Y, X is called a *parent* of Y. Y is called a *descendant* of X just in case there is a directed path from X to Y consisting of arrows lining up tip-to-tail to link intermediate variables. By convention, every variable is a descendant of itself. If Y is a descendant of X, X is an *ancestor* of Y. A DAG is *acyclic* insofar as it contains no 'loops', i.e. no variable is an ancestor of itself. We will call $\mathbf{PA}(X)$ the set of parents of X, and $\mathbf{DE}(X)$ the set of descendants of X.

Not all DAGs are Bayes nets. To count as a Bayes net, a DAG **G** on a set of variables **V** must satisfy two conditions that connect it to the probability distribution over **V**. The first one is the *Markov condition*.

> **Markov Condition (MC)**. Take any variable X in **V**, and let $\mathbf{V}\backslash\mathbf{DE}(X)$ be the set of variables in **V** that are not descendants of X. Then for any set of variables **Y** in $\mathbf{V}\backslash\mathbf{DE}(X)$, $P(\mathbf{Y}/X.\mathbf{PA}(X)) = P(\mathbf{Y}/\mathbf{PA}(X))$.

In words: a graph satisfies the Markov condition just in case for any variable X, the set of parents of X screens off X from its non-descendants. Thus, when a graph satisfies **MC** we can extract information about conditional independence relationships from the graph: it is in that sense that the graph (partially) represents the probability function. To illustrate the Markov condition, suppose you have a set of variables $\{X,\ Y\}$ such that X and Y are probabilistically dependent. Then the graphs in Figures 2.4 and 2.5 satisfy the Markov condition, but the graph in Figure 2.6 doesn't. Since in Figure 2.6, Y is a non-descendant of X and there are no parents of X, **MC** would imply that Y is independent of X, contrary to fact.

In addition, to be a Bayes net a DAG must satisfy the *Minimality condition*. Say that a graph **G'** is a *proper subgraph* of **G** if **G** and **G'** are defined over the same set of variables, all the arrows in **G'** are in **G**, and some of the arrows in **G** are not in **G'**.

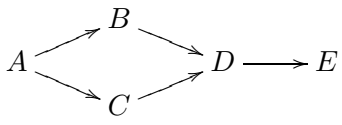

Figure 2.7

The minimality condition says the following:

Minimality Condition (Min). No proper subgraph of **G** satisfies **MC**.

The idea is that a graph satisfies **Min** when it contains the least number of edges between variables consistent with the **MC**. To illustrate, suppose that X and Y are independent. Then the graphs in Figure 2.4 and 2.5 satisfy **MC** but not **Min**, as their subgraph in Figure 2.6 also satisfies **MC**.

To summarize, a Bayes net over a set of variables is a convenient way to represent certain dependence and independence relationships between the variables. For instance[38], suppose that A represents the season of a given year, B the rain fall during the season, C whether the sprinkler is on during the season, D whether the pavement is wet and E whether the pavement is slippery. The probabilistic relations between these variables can plausibly be represented by the Bayes net in Figure 2.7. This graph represents various independence relationships between the variables. For instance, it says that whether the pavement is wet is independent of the season, if one holds fixed the rain fall during the season and whether the sprinkler is turned on. It also says that whether the pavement is slippery is independent of all other variables if one holds fixed whether the pavement is wet; and so on.

Before we move on, let me make two important remarks. First, note that a Bayes net may not represent *all* the dependence and independence relationships between variables. For instance, the graph in Figure 2.7 doesn't allow us to determine whether A and D are independent or not: no information about the probabilistic relations between the two variables can be extracted from **MC**. The second, crucial remark is that several Bayes nets may be compatible with the probability distribution over **V**. For instance, suppose that we know that X and Z are dependent unconditionally and independent

[38]This is a standard example that can be found in Pearl (2009) and elsewhere.

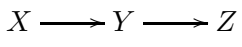

Figure 2.8

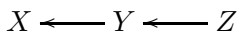

Figure 2.9

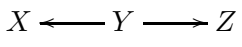

Figure 2.10

given Y. Then the graphs in Figures 2.8, 2.9 and 2.10 are all Bayes nets on $\{X, Y, Z\}$.

Moving now to causal structure, note that the causal relations between variables can *also* be represented by a DAG if we interpret the arrows as representing causal relations. A DAG that represents causal structure is called a *causal graph* (Spirtes et al., 2000, 24). More precisely, let's say that a DAG $\mathbf{G}$ on $\mathbf{V}$ is a causal graph on $\mathbf{V}$ just in case the following condition is satisfied: $\mathbf{G}$ contains an arrow from X to Y iff X is a *direct cause* of Y, i.e. a cause of Y whose influence on Y isn't mediated by any of the other variables in $\mathbf{V}$. In a causal graph, the set of parents of a variable X is the set of its direct causes $\mathbf{DC}(X)$. Consider for instance the graph in Figure 2.7. As we have seen, this graph can be used to represent the *probabilistic* relations between the relevant variables. But the graph can also be interpreted as a causal graph representing the *causal* relations between these variables. Thus interpreted, the graph says that the rain fall is a direct cause of the pavement being wet, that the pavement being wet is a direct cause of its being slippery, and so on.

So far we haven't said anything about the relations between probabilistic and causal structure; we have simply introduced two graph-theoretic tools (Bayes nets and causal graphs) to represent each. But the core idea of probabilistic theories of causation - and the central idea of SGS - is that there is a close connection between causal and probabilistic dependence, so that the two kinds of representation coincide. More formally: SGS posit that for any X and Y, there is a set of variables $\mathbf{V}$ that includes X and Y, such that the correct causal graph on $\mathbf{V}$ is a Bayes net. If $\mathbf{V}$ is a 'suitable' set of variables (more on this in a minute), then the correct graph $\mathbf{G}$ of the causal relations between the variables in $\mathbf{V}$ obeys **MC** and **Min**. Since Bayes nets represent probabilistic structure, this is to assume that if the variables in $\mathbf{V}$ are causally related in certain ways, they must also be probabilistically related in a certain fashion. A bit more precisely, SGS's claim amount to the assumption that if $\mathbf{V}$ is a suitable set of variables

and **G** the correct causal graph on **V**, the relations of direct causation represented by the arrows in **G** satisfy the following two conditions:

> **Causal Markov Condition (MC)**. Take any variable X in **V**, and let **V****DE**(X) be the set of variables in **V** that are not descendants of X. **DC**(X) is the set of direct causes of X. Then for any set of variables **Y** in **V****DE**(X), $P(\mathbf{Y}/X.\mathbf{DC}(X)) = P(\mathbf{Y}/\mathbf{DC}(X))$.
>
> **Causal Minimality Condition (CMin)**. No proper subgraph of **G** satisfies **CMC**.

Each condition is obtained by interpreting parenthood relations in **MC** and **Min** as relations of direct causation. Of these two conditions, the causal Markov condition is the most important one, as it embodies a substantial assumption about how causation relates to probabilities.[39] **CMC** says that conditional on its direct causes, X is independent of every other variable in **V**, except for its effects. For instance, applied to the graph in Figure 2.7 (interpreted as a causal graph), **CMC** says for instance that given the rain fall and the sprinkler being on (or off), whether the pavement is wet is independent of the season; that given that the pavement is wet, whether it is slippery is independent of the season, the amount of rain fall, and the status of the sprinkler, and so on. **CMC** is closely connected to Reichenbach's principle of the common cause; indeed, it entails the latter. To see this, note that if X and Y are not related as cause and effect and have no ancestors in common, **CMC** entails that they are probabilistically independent conditional on the empty set - i.e. unconditionally independent. Conversely, if X and Y are correlated, **CMC** entails that either one causes the other or they have common ancestors (i.e. common causes), and that in the latter case, these common causes screen off the correlation. This is Reichenbach's principle of the common cause.

As I mentioned, SGS do not assume that *any* causal graph on a set of variables **V** is a Bayes net; in addition, **V** must be a 'suitable' set of variables. To illustrate, suppose that C and D are common causes of two independent effects A and B, as represented in Figure 2.11. And now suppose that our set of variables is {C, A, B}. The correct

[39] By contrast, the causal minimality condition is more a convention about representing the relations between causation and probabilities than a substantial assumption about those relations.

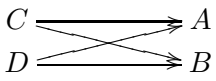

Figure 2.11

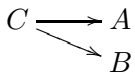

Figure 2.12

causal graph is the one in Figure 2.12, but this graph isn't a Bayes net as it doesn't satisfy **MC**. Conditional on C, A and B are still correlated. Thus, SGS say that to be 'suitable', a set of variables **V** should be 'causally sufficient': that is, all the common causes of two variables in **V** should be included in **V**. It is only when **V** is causally sufficient that the causal graph (SGS say) will also be a Bayes net.

Several objections have been raised against the **CMC**. First, it is often claimed that the CMC fails for quantum systems involving distant correlations (see e.g. Papineau (1993), Spirtes et al. (2000)). For instance, there are correlations between spacelike separated measurements of (e.g.) the vertical spin of particles in the singlet state. For each of them, the probability of spin up is $\frac{1}{2}$, but the probability of both of them having spin up is not $\frac{1}{4}$ but $\frac{1}{2}$, which means that the results of the two measurements are correlated. And it can be shown that there cannot be any *local* common cause that screens off the results of the measurements from each other. Thus, it is concluded, there is no set of variables representing the system in which the CMC holds. But this argument relies on dubious assumptions about the relation between locality (in the relativistic sense of the term) and causation. In particular, the argument assumes that (a) since the results of the measurements are spacelike separated they cannot be causally related, because influence cannot travel from one to the other and (b) any common cause of the results must be a *local* one - i.e. one that is related to the two effects through a local process. But if causation is fundamentally a matter of probabilistic connections rather than (say) physical processes, these requirements appear arbitrary, and there is no reason to exclude the possibility of non-local causation at the outset.

Salmon (1980) offers another kind of putative counterexample to the CMC. He imagines a novice billiard player who has a $\frac{1}{2}$ chance of sinking the eight ball. Suppose that the case is such that for all the ways in which the eight ball might sink, the cue

ball will almost certainly sink as well. Then the eight ball sinking is correlated with the cue ball sinking, but the only apparent candidate for the status of common cause - the cue ball being struck - doesn't screen off the correlation. To put the point epistemically, if you know that the cue ball has been struck, learning in addition that the eight ball sank gives you additional evidence that the cue ball sank as well. But Spirtes et al. (2000) plausibly respond that if we specify the details of the way in which the cue ball was struck (in particular the momentum it imparts to the ball, among other things), we can regain conditional independence. So as long as the variable representing the striking of the cue ball is fine-grained enough the relevant graph will obey the CMC.

A third putative counterexample to the CMC is offered by Sober (1988, 2001). Sober notes that both bread prices in London and sea levels in Venice have been steadily increasing over the last two centuries. Thus, he says, there is a correlation between the two. Given a high bread price, the sea level in Venice is more likely to be high as well. But there is no causal connection between the two trends, so that the correlation cannot be accounted for along the lines of CMC. The problem with this argument is that it isn't clear that there really is a correlation between the bread price at t and the sea levels at t. However, sampling bread prices and sea levels at times other than t are not reliable ways to establish a correlation between these two events. The reason is that the distribution from which we are sampling isn't a stationary one. (Roughly, a frequency distribution is stationary when it is invariant with respect to time.) And sampling from non-stationary distribution is not a reliable statistical method to establish a correlation at a time, as the following example shows.[40] Consider a population of six-year-old students. Presumably inside this population there is no correlation between height and mathematical ability. But if we take samples from the population at *later* times, we will find that taller heights are associated with better mathematical abilities (as both quantities grow over time). I conclude, then, that the **CMC** is a relatively secure assumption: counterexamples to it dissolve upon closer examination. Its truth will be assumed in what follows.

The question with which SGS are concerned, remember, is under which conditions

[40] I borrow this example from Hoover (2003).

the causal relations between variables can be inferred from information about the probability distribution over this set of variables. Note that the claim that causal graphs are Bayes nets doesn't give us a one-to-one correspondence between causal and probabilistic structure. As we have seen, several Bayes nets may be compatible with a single probability distribution, so that information about the probability distribution by itself doesn't select a unique causal hypothesis. For instance, if all we know is that X and Z are correlated unconditionally and uncorrelated given Y, we cannot distinguish between the three nets in Figs. 2.8, 2.9 and 2.10. That is, Y may be a common cause of X and Z or a causal intermediary between X and Z. But SGS show that if we make a further assumption about the relation between causation and probabilities, the space of possible causal orderings that might obtain between variables becomes significantly constrained. This further assumption is that the correct causal graph **G** on a set of variables **V** satisfies the *Faithfulness Condition*:

> **Faithfulness Condition (FC)**. All the (conditional or unconditional) probabilistic independence relations among the variables in **V** are entailed by **CMC**.

To illustrate, consider the graph in Figure 2.7 and assume that A and D are independent of each other. Then if interpreted as a causal graph the graph doesn't satisfy **FC** since this independence isn't entailed by **CMC**. (Colloquially, the graph isn't faithful to the probability distribution.) **FC** prohibits the following kind of situation.[41] Suppose that living in the countryside (C=1) rather than in the city (C=0) causes one to smoke (S=1) and thereby to have lung cancer (L=0). And suppose in addition that living in the countryside also has a negative causal influence on lung cancer (perhaps by making you less likely to be exposed to certain carcinogenics). The causal structure is represented in Figure 2.13. Now suppose that the parameter values are such that the positive causal influence of S on L is canceled by its negative probabilistic relation to L due to C. Then we will find that S and L are uncorrelated. Since this independence isn't entailed by the **CMC**, the graph is unfaithful to the distribution: the independence isn't due to the causal structure, but to the parameter values. This example also shows

[41] I borrow this example from Hitchcock (2012), with some modifications.

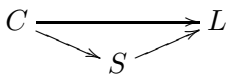

Figure 2.13

that the assumption that every (true) causal graph satisfies **FC** is problematic, since the case just described seems perfectly possible. This is arguably not a problem for SGS given the methodological nature of their project. SGS are concerned with building reliable discovery algorithms to extract causal information from statistical data. As long as unfaithful causal structures are rare in our world, discovery algorithms that presuppose **FC** are still reliable. And SGS indeed offer a convincing measure-theoretic argument to the effect that unfaithful causal structures are at most exceptional in our world (2000, Theorem 3.2, 41-42). As we will see later **FC** is much more problematic if we try to use SGS's framework for metaphysical purposes.

FC allows us to establish two very significant results. The first one is due to Pearl and Verma (1991). Suppose you have three variables satisfying the following conditions: X and Y are both correlated with Z, X and Y are independent, but dependent given Z. In that case Z is said to be an *unshielded collider*. Then the only Bayes net that satisfies **FC** is the one in Figure 2.14. In other words, the only causal hypothesis compatible with the probabilistic facts is that Z is a common effect of X and Y. This gives us an important and precisely formulated asymmetry between causes and effects. When two effects of a (single) common cause are causally independent (i.e. neither causes the other), then the two effects are unconditionally correlated but independent given the common cause. When two causes have a common effect and are not causally connected to each other, this is the other way around. The two causes are unconditionally independent but correlated given their common effect.[42]

The second result is established by SGS in their Theorem 4.6 (2000, 65). Say that two causal graphs on a set of variables **V** are statistically indistinguishable just in case they are both consistent with the probability distribution over **V**. SGS show that for

[42]Here is an example from Hausman (1998) to illustrate the latter claim. Suppose that whether the light is on is a function of two switches: both must be in the same position for the light to be on. Then the positions of the switches are uncorrelated with each other, but correlated given that the light is on. Given that the light is on, the positions of the two switches are almost perfectly correlated.

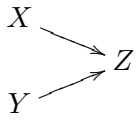

Figure 2.14

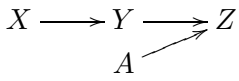

Figure 2.15

any pair of statistically indistinguishable graphs on **V**, there is a wider set of variables **V'** and a possible probability distribution over **V'** consistent with the actual probability distribution over **V** such that there is a unique causal graph on **V'**. In other words, there is always in principle a wider set of variables and probability distribution over them that distinguishes between the competing causal hypotheses. SGS's proof of this result exploits the phenomenon of unshielded colliders. It can be illustrated as follows in the case of the three statistically indistinguishable graphs of Figs. 2.8-2.10. Suppose that there is a variable A that is unconditionally independent of X, correlated with Y and Z, and dependent on X given Y. Here Y is an unshielded collider for X and A. Then the only Bayes net on $\{X, Y, Z, A\}$ that satisfies the **FC** is the graph in Figure 2.15. in which Y is a common effect of A and X. So here the information about A and its probabilistic relations to other variables allows us to determine the causal structure between X, Y and Z: Y is a causal intermediary on a causal path from X to Z. This theorem is the crucial result used by Papineau to develop a reductive metaphysical theory of causal dependence.

3.2.2 Papineau's Theory

Theorem 4.6 only says that in cases of statistical indistinguishability there is a *possible* probability distribution over a wider set of variables that uniquely determines the causal structure. Papineau goes further and makes the assumption that for any two statistically indistinguishable causal graphs over a set of variables, there is always a wider set of variables such that the *actual* probability distribution over it uniquely determines causal structure. We will examine in the next subsection whether Papineau has the right to avail himself of this assumption. For now, note that if this assumption

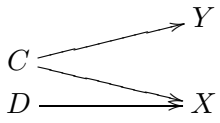

Figure 2.16

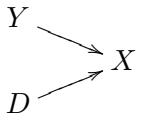

Figure 2.17

is correct one can obtain a reduction of causal structure to probabilistic structure, as follows. Take two events c and e represented with variables A and B respectively. (A=1 represents c's occurrence, A=0 c's non-occurrence; similarly for B.) From Theorem 4.6 and Papineau's assumption, it follows that there is a set of variables $\mathbf{V}$ such that the actual probability distribution over $\mathbf{V}$ uniquely determines a single causal graph $\mathbf{G}$ over $\mathbf{V}$: the probabilistic structure determines the causal structure. We can then determine whether c causes e by checking whether there is a directed path from A to B in $\mathbf{G}$.

There are two complications here. First, remember that SGS's results hold only for *causally sufficient* sets of variables (sets that include all common causes of variables in the set). Without this assumption Papineau's account runs into troubles. For instance, suppose that we have a causal structure as represented in Figure 2.16. And now suppose that our set of variables omits C. The only causal graph on $\{X,\ Y,\ D\}$ is the one represented in Figure 2.17: there X is mistakenly counted as an effect of Y. This mistaken judgment will be reverted only if the common cause C is included in the set. So it is crucial for Papineau to find a way to require sets of variables to be causally sufficient; but for the theory to be reductive Papineau cannot make use of the very notion of causal sufficiency. One suggestion is to add a requirement of *stability* to the theory: c should be counted as a cause of e only if there is a set of variables over which the unique causal graph contains a directed path from c to e, and there is no wider set of variables in which this verdict is overturned. Since a set containing C will overturn the misleading verdict of Fig 2.17, this requirement entails that X isn't an effect of Y after all.

A second complication is that the existence of a directed path from C to E isn't a

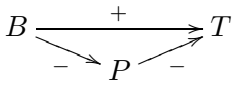

Figure 2.18

sufficient condition for E to causally dependent on C. To see why, consider the following example due to Hesslow (1976). Taking birth control pills has two causal effects on thrombosis in the population of women who are fertile, sexually active and under 35. On the one hand, birth control pills prevent thrombosis, since they prevent pregnancy, which is itself a major causal factor for thrombosis. On the other hand, birth control pills also have a positive causal effect on thrombosis.[43] The causal structure is represented in Figure 2.18, with B for the pills, P for pregnancy and T for thrombosis. Now suppose that the negative causal route is much stronger than the positive one, so that on the whole women who take birth control pills are much less at risk of thrombosis than others. In that case we would like to say that thrombosis is negatively causally dependent on taking the pill. (An agent whose sole goal is to avoid thrombosis and who has the choice of taking birth control pills should choose to do so, as long as she belongs to the relevant population.) But any correct causal graph over a set of variables including B, P and T will include a positive directed path from B to T. Here is a solution. To check whether E negatively causally depends on C, one should hold fixed all direct causes of E such that there is no directed path from C to them. E negatively causally depends on C just in case it is negatively correlated with E holding fixed this set of direct causes. (The same holds *mutatis mutandis* for positive causal dependence, of course.) By not holding fixed those direct causes such that there is a directed path from C to them, one is sure to take into account all the causal routes from C to E. Applied to Hesslow's case, this procedure tells us correctly that thrombosis negatively causally depends on birth control pills, since when we do not hold P fixed T and B are negatively correlated.

Putting all of this together, we arrive at the following theory:

[43] This is established by the fact that among women who do not fall pregnant, those who take birth control pills are more at risk of having thrombosis than others.

> **Papineau)**. Take two events c and e represented with variables C and E respectively. Then c is a positive difference-maker for e just in case there is a unique graph **G** over a set of variables **V** including C and E such that
>
> 1. Holding fixed all direct causes of E such that there is no directed path from C to them, E=1 is positively correlated with C=1
> 2. There is no graph *G'* over a wider set of variables **V'** in which condition (1) doesn't hold.

(2 is the aforementioned stability condition.) Papineau's theory doesn't make unconditional correlation necessary for causation. A cause need only be correlated to its effect e conditional on a certain set of direct causes of e. The theory thereby correctly handles Simpson's paradox case of Fig. 2.3. which raised trouble for Reichenbach's account. Although smoking is negatively correlated with heart attacks, the positive correlation is restored once one conditionalizes on the other direct cause of heart attacks (exercising).

Papineau's theory gives us an attractive explanation of the causal direction and time-asymmetry. On his view, the causal direction is at bottom a product of the following phenomenon. Suppose that we have a causal correlation between X and Y. Then in general one will find another variable D that is correlated with Y, uncorrelated with X unconditionally but dependent on X given Y, as in Fig. 2.19. Here Y is an unshielded collider for X and D. (One should think of D here as another cause of Y.) By contrast, if X is an effect of Y, it is the other way around: in general, if D is correlated with Y it will also be correlated with X and screened-off from X by Y. See Fig. 2.20, where the dotted line indicates the presence of a correlation between D and X.[44] Like in Reichenbach's theory, cause and effect can be distinguished by their probabilistic relations to each other *and to other events*: the causal direction emerges from asymmetric features of the probabilistic network of events in our world.

On Papineau's view, the fact that causes typically precede their effects is a product of a pervasive statistical asymmetry that is sufficiently close to what Reichenbach called the *fork asymmetry* to deserve the name as well. It consists in the following fact.

[44] Note that these statements hold only 'in general', because the relevant probabilistic patterns may be masked by other causal relations between the variables. For instance, if X causes Y but also causes D, one will find a correlation between X and D. In those more complicated cases, SGS's rules for constructing causal graphs will still allow us to determine that X causes Y if given probabilistic information.

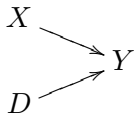

Figure 2.19

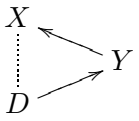

Figure 2.20

In general, when two variables A and B are both correlated with a third variable C in their common past, A and B are correlated with each other. (In those cases A, B and C form a fork open to the future.) For instance, lung cancer and yellow finger are both correlated with smoking earlier, and are themselves correlated. By contrast, when two variables A and B are correlated with a third variable C in their common *future*, they are *not* thereby correlated with each other. For instance, smoking and working in an asbestos factory are both correlated with lung cancer later, but uncorrelated with each other. The relevant asymmetry, then, is the fact that an event typically induces correlations between *later* events with which it is correlated, while it doesn't induce correlations between *earlier* events with which it is correlated. Now suppose that we have two (independent) causes X and D of a variable Y, as in Figure 2.19. On Papineau's view what makes X and D causes of Y is the fact that they are both correlated with Y but independent of each other. But then in virtue of the fork-asymmetry, this means that D and X must be temporally earlier than Y. If both D and X were later than Y, the prevalence of forks open to the future would ensure a correlation between them. Contrast this with a case where Y is a common cause of X and D, as in Figure 2.21. On Papineau's view what makes them effects of Y is that X and D are both correlated with Y and correlated with each other (although the correlation disappears given Y). In turn, given the fork asymmetry this means that X and D must be temporally later than Y. If both X and D were earlier than Y, their correlation with Y would not induce a correlation between them. By contrast to Reichenbach's account, this explanation doesn't fix the temporal direction of causation via the *global* direction of open forks. Rather, the temporal direction of a causal process is fixed by the local forks in which it is embedded. It thus leaves a

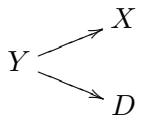

Figure 2.21

loophole for backward causation in exceptional circumstances in which forks run counter to their usual direction.

3.2.3 Problems for Papineau's Theory

Papineau's theory is a very powerful account of causal dependence, and my own account in chapter 4 will make use of very similar ideas. As it stands, however, the account suffers from two problems.

The first problem is that Papineau's theory relies on two very controversial assumptions. The first one is the assumption that for any causally ambiguous probabilistic structure over a set of variables, there is not only a possible (as per SGS's Theorem 4.6) but an *actual* wider probabilistic structure that can disambiguate it. The problem here is that this assumption needs to be defended, but Papineau says very little to motivate it. The only defense he offers goes as follows. It may well be conceptually possible that X causes Y but that no amount of information about actual probabilistic structure allows us to conclusively establish this. But conceptual possibility doesn't entail metaphysical possibility, and one can coherently maintain that such cases are not in fact metaphysically possible. As Hausman says, 'those who do not already accept [Papineau's] reduction may not find this response persuasive' (1998, 221). The second controversial assumption is that every actual causal structure satisfies the faithfulness condition. Although **FC** is fine as a methodological assumption, it is far less plausible as a metaphysical one. The problem here is that realistic unfaithful causal structures seem perfectly possible, and for all we know our world might contain some.[45] One such realistic case is Hitchcock's countryside/smoking example presented earlier (see Fig. 2.13). In this case, Papineau wrongly entails that smoking isn't a cause of lung cancer,

[45] Although for reasons pointed out by SGS we should expect such cases to be rare.

hence that refraining to smoke isn't an effective strategy to avoid the latter.[46]

The second issue is that, like other theories we have seen so far, Papineau's account leaves it mysterious why difference-making as it construes it is the relation that matters for rational decision-making. In fairness to Papineau, it should be noted that he clearly recognizes that there is a challenge there and tries to address it (see Papineau (1993, 246-50)). In his case the challenge takes the following form. As Papineau points out, his reduction relies *inter alia* on the idea that cause and effect are correlated, and the idea that influence is a matter of correlation is in itself very plausible. But of course one needs to explain why only certain sorts of correlations can be exploited (i.e. genuinely causal rather than spurious correlations), and only in the direction from cause to effect. So the task for Papineau is to show that the features that distinguish causal from spurious correlations and fix the causal direction on his view also explain why only certain correlations can be used to advance our goals.

Papineau argues that for c to be a good way to advance a goal e, correlation isn't sufficient. The correlation should be *robust* under variations in which c is brought about. That is, the correlation should remain stable if one conditions on a direct cause of c (in the graph-theoretic sense of 'direct cause'). This amounts to the requirement that for any direct cause d of c, c screens off d from e: $P(e/c) = P(e/c.d)$. Now consider a case where an action and a desired outcome stands in a spurious correlation (a hypothetical example that, by the way, we will examine in detail in the next chapter):

> *Chocolate.* Suppose that people who eat chocolate regularly tend to live longer. Eating chocolate, however, is causally irrelevant to life expectancy. Instead, the correlation is due to a common cause: the presence of a certain gene which causes one both to eat chocolate and (via a different causal route) to live longer. The causal structure of the case is represented in Figure 2.22.

In this case the correlation between eating chocolate and life expectancy isn't robust: in fact, the correlation simply *disappears* if one holds fixed the presence or absence of the (direct) cause of eating chocolate, viz. the gene. On Papineau's view, this is why

[46] Papineau (1993) argues that the faithfulness assumption is in fact dispensable. His argument has been convincingly refuted by Hausman (1998, 219-20), as Papineau (2001b) himself recognizes.

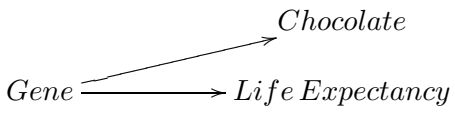

Figure 2.22

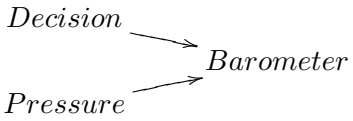

Figure 2.23

the correlation and similar spurious correlations cannot be exploited. Now consider a genuinely causal correlation such as the correlation between atmospheric pressure and the reading on a barometer. And suppose that the reading on the barometer is partially controllable by human decision (Figure 2.23). In the special case where the decision and the atmospheric pressure are independent of each other, we should have $P(Pressure/Barometer.Decision) \neq P(Pressure/Barometer)$.[47] This means that the barometer-pressure correlation isn't robust under the direct causes of the barometer reading (in particular the decision to manipulate it). According to Papineau this explains why the barometer-pressure correlation cannot be exploited so as to influence atmospheric pressure, and more generally why causal correlations can only be exploited in the cause-to-effect direction.

There are two problems with this explanation, however. First, it isn't clear why we should only care about *robust* correlations between act and desired outcome in the context of rational decision-making. Papineau defends this assumption as follows:

> The link between causation and decision requires more than just that a cause makes *some* probabilistic difference to its effect. [W]e are interested in [causation] because knowledge of causal connections enables us to choose means appropriate to our ends. Such decisions, however, are characteristically quantitative. We want to know how likely it is that E will follow C, so as to be able to compare the overall advantage expected from C with those from other courses of action. But this means that a C/E link that had different strengths in different circumstances would not qualify as a causal connection. Just knowing that C makes some probabilistic difference to E is unhelpful in most real-life decisions. We need to know how much difference it makes. (1993, 250)

But that doesn't seem right. Even if we must know the exact value of a correlation to exploit it, correlations that vary in strength depending on the circumstances may still

[47] As we have seen, independent causes of a joint effect are correlated given the effect. So in this case there should be a correlation between *Decision* and *Pressure* given *Barometer*; but this is just to say that *Barometer* doesn't screen off *Pressure* from *Decision*. Papineau (1993, 247-8) argues that this is also true in more complicated situations where pressure and decision are causally related.

be exploitable if we know what circumstance we are in. For instance, in the barometer case, if I know that the barometer reading will come about partially as a result of my decision and I know the value of the barometer-pressure correlation given my decision, I will be able to know exactly which probabilistic difference manipulating the barometer will make to atmospheric pressure. Another problem with Papineau's explanation is that there are intuitively exploitable correlations between an act and an outcome that are not stable under the direct causes of the act. Consider again Cartwright's example in which smoking is a cause of heart attacks and exercising both causes smoking and prevents heart attacks (see Figure 2.3). Intuitively in that case it is an effective strategy to refrain from smoking in order to avoid heart attacks, but the correlation between smoking and heart attacks isn't stable under the direct causes of smoking: conditioning on exercising changes the value of the correlation.

Chapter 3

An Evidentialist Theory of Exploitable Correlations

We saw in the last chapter that none of the current attempts to answer Russell's problem manage to provide an entirely satisfactory solution. In particular, all of them make it in some way or other mysterious why difference-making is the sort of relation that matters for rational decision-making. In other words, they all have difficulties explaining the practical relevance of causation. Our review of existing solutions to Russell's problem has also yielded more positive results, however. More specifically, we have seen that causation is plausibly taken as in part a matter of *correlation* or *probabilistic dependence* between the cause and the effect (where the relevant probabilities are statistical-mechanical Humean chances). There are three reasons that make this idea attractive. First, considerations related to the Albert-Elga objection to Lewis seem to show quite decisively that any plausible account of e's causal dependence on c will have to appeal to facts about the conditional probabilities of e given c and $\sim c$. Second, this solves the problem of localization. Although the nature of physical laws prevent localized events to be *physical determinants*, it doesn't prevent localized events to be *statistically associated* with others. Third, there is something self-evident to the idea that goal advancement is at bottom a matter of correlation between action and desired outcome, so that a probabilistic theory of causal dependence stands in a good position

to explain the practical relevance of causation.

Nevertheless, there are two challenges for a theory that takes as a starting point the idea that causation is a matter of correlation. First, it must provide an extensionally adequate account that correctly distinguishes spurious from genuinely causal correlations and explains where the causal direction and time-asymmetry come from.[1] Second it must explain why, if influence is fundamentally a matter of correlation, only certain correlations are exploitable and in one direction only. As we have seen, the most developed attempt to solve these two challenges (due to Papineau) doesn't entirely succeed. My goal in this chapter and the next is to provide a better solution to these two problems.

My strategy will be to deal with the latter problem first by offering a plausible theory of what makes a correlation exploitable for advancing goals. To articulate it I will use a decision theory called *evidential decision theory* (EDT). EDT was first proposed by Jeffrey (1983).[2] In this chapter, I will argue that correlations that can be exploited are exactly those that are good for acting on according to EDT. Indeed, I will argue that EDT provides a good explanation of why a correlation between an act and an outcome cannot be exploited when the act doesn't cause the outcome. In fact, a leitmotiv of this chapter and the next will be that there is a very close connection between the notion of a causal correlation and the notion of a correlation exploitable according to EDT. (More on the nature of this connection in a minute.) What makes this approach attractive is that EDT's guiding idea is precisely that goal advancement is a matter of correlation - i.e., that an act being a good way to achieve an outcome is a matter of the act being correlated with the outcome. Since there is something very plausible about this conception of influence, a theory of causation that relies on EDT to explain why only causal relations can be usefully exploited is guaranteed not to make the practical relevance of causation a mystery.

Nevertheless, the idea that EDT can be the basis of a plausible theory of exploitable

[1] It must also handle the complication raised by cases involving Simpson's paradox, which show that a correlation between a cause and its effect may appear only when certain factors are held fixed.

[2] Eells (1981, 1982, 1984), Horwich (1985, 1987), Price (1986, 1991, 1992, 2012) and Ahmed (2010) also offer important defenses of EDT.

correlations may seem crazy. The main and seemingly devastating objection to EDT is that it recommends doing an act for the sake of an outcome *whenever* act and outcome are correlated. EDT thus seems to entail that *any* act-outcome correlation whatsoever can be exploited (including spurious ones), and that a causal correlation can be exploited in in the effect-to-cause direction. I will argue that EDT in fact doesn't have these consequences. Together with independently plausible principles of rational choice, it entails that only those correlations that are intuitively causal can be exploited, and only in the cause-to-effect direction.

In chapter 4, I will use the theory of exploitable correlations developed in the present chapter to give a new account of causal dependence. It will be useful to give a brief preview of how this will go. To a first approximation, the idea is that we can simply *identify* causal correlations with those that are exploitable according to EDT: what makes a correlation causal is that it can be exploited according to EDT. However, this isn't entirely satisfactory. According to EDT, only correlations between *human actions* and other events can be exploited. We cannot exploit the causal correlation between the position of the moon and the tides because the position of the moon isn't the sort of thing we can deliberate about. A proper theory of causation should handle causal relations that involve unmanipulable causes. My strategy to do so will be similar to the one adopted by *interventionists*.[3] Like me, interventionists are concerned with offering a theory of causation that makes good sense of its practical relevance. (However, they are not looking for a reductive theory.) They start with the idea that a correlation between e and c can be exploited when the correlation survives under a human *manipulation* of c. That is, the correlation is exploitable iff when I manipulate or 'wiggle' c, e is more likely to happen. For instance, the barometer cannot be used to influence atmospheric pressure because when I manipulate the reading on the barometer, the correlation between the reading and atmospheric pressure *disappears*, as the atmospheric pressure has no influence on the barometer reading anymore. To turn this idea into a theory of causation that handles unmanipulable causes, they appeal to the idea that causal correlations (whether exploitable or not) are those that survive a process *relevantly like*

[3]See for instance Meek and Glymour (1994), Pearl (2009) and Woodward (2003).

a human manipulation called an *intervention*. The idea is that an intervention should have all the characteristics that human manipulations have and that explain why a correlation can be exploited only when it is causal. In the example I just gave, the crucial property of a manipulation of the barometer is that the manipulation completely controls the setting of the barometer, so that the correlation between barometer and pressure disappears given it. It is this feature of human manipulation that explains why the relation between pressure and barometer cannot be exploited in the effect-to-cause direction. An intervention on the barometer should have the same characteristic: it should be a process that completely causally controls the reading of the barometer. More generally, an intervention on an event c should be a 'surgical' causal process that entirely controls c, so that it breaks the connections between c and its normal causes.[4] By requiring causal correlations to survive under interventions, interventionists can explain the practical relevance of those correlations on the basis of the fact that manipulations are interventions. But an intervention need not be a human action. For instance an intervention on the barometer reading might be the operation of a mechanical device that randomly sets the barometer dial at various positions. Thus in principle unmanipulable causes can be intervened upon.

Now, the interventionist account of causal dependence is an explicitly non-reductive one. The notion of intervention is spelled out in causal terms. So interventionism doesn't offer a solution to Russell's problem.[5] But the interventionist *strategy* can be used for reductive purposes. I will argue in this chapter that for a correlation between an act c and an event e to be exploitable, c and e must stand in certain probabilistic relations to a third event, the *deliberation* of an agent trying to decide whether to do c. In chapter 4, I will apply the interventionist strategy and argue that causal correlations are those correlations that stand in certain probabilistic relations to a third event that is *relevantly like* deliberation. I will call such a third event a *probabilistic intervention* or *p-intervention* for short. The idea is that a p-intervention must have the same features that deliberation has and which explain why only certain correlations can be

[4]See Woodward (2003, 94-102) for a much more precise formal characterization of intervention.
[5]This is of course not to say that it does not illuminate certain crucial facts about causation.

exploited. But a p-intervention need not be a deliberation or anything agentive, so that in principle p-interventions can occur on unmanipulable causes. The resulting theory explains why only causal relations can be exploited, as what makes a correlation exploitable is precisely its having a certain relation to an event relevantly like a p-intervention (namely deliberation). But the theory also handles causes which are not human actions, since a p-intervention can occur on an event that is not a human action. By contrast to the notion of intervention, the notion of a p-intervention will be spelled out entirely in non-causal terms, so that the resulting theory is reductive.

To summarize, then, I have two goals in this chapter. The first is to show that EDT provides a plausible theory of exploitable correlations; in particular, that it entails that an act can be exploited to influence an outcome only when the act is a cause of the outcome. My second goal is to identify those features of deliberation that make a correlation exploitable or not according to EDT. This will allow me to define the notion of a p-intervention in chapter 4.

A last preliminary remark: I am not the first one to defend the idea that EDT gives us a good explanation of exploitable correlations, and that we can build a plausible solution to Russell's problem on its basis. This view has also been defended by Price (1992; 1996; 2007; 2012), partly in collaboration with Brad Weslake (Price and Weslake, 2009).[6] But I will develop this idea in a way substantially different from Price. Price's account relies crucially on a defense of EDT known as the *tickle defense*, which attempts to show that spurious correlations are not good for acting on because those correlations *disappear* conditional on the evidence possessed by a deliberating. I will argue that Price's use of the tickle defense in his theory of exploitable correlations raises severe difficulties for his view. I will propose a different defense of EDT which makes use of certain elements of the tickle defense but escapes these difficulties. The guiding idea of this new defense that spurious correlations between an act and an outcome cannot be exploited because they do not induce correlations between *deliberating* about whether to perform the act on the one hand and the occurrence of the desired outcome on the

[6] Meek and Glymour (1994) and Hitchcock (1996a) also argue that the practical relevance of causation can be explained through EDT, although they do not try to derive a solution to Russell's problem from this explanation.

other hand.

§1 starts with some background on EDT and the sort of decision situation on which I focus in this chapter. §2 presents the tickle defense, and §3 criticizes Price's attempt to use the tickle defense to give an evidentialist theory of exploitable correlations. §4 presents a new defense of EDT and extracts a general theory of exploitable correlations on its basis. For the purposes of chapter 4 it will be important to characterize the relation that a suitable deliberating agent must have to her contemplated actions in non-causal terms. This is the topic of §5.

1 Background

Let me start with some background about decision theory. Decision theory gives recommendations about what to do in a decision situation - a situation where an agent has to make a choice between several possible actions. A decision situation can be represented as a quadruple $\{\mathfrak{A}, \mathfrak{O}, C_t, U\}$. $\mathfrak{A}$ is the set of alternative actions available to the agent: her *options*. By convention, the elements of $\mathfrak{A}$ form a partition: they are mutually exclusive and collectively exhaustive (so that the agent must choose one and only one option in $\mathfrak{A}$). $\mathfrak{O}$ is the set of outcomes that might accompany the options. Like the elements of $\mathfrak{A}$, the elements of $\mathfrak{O}$ are assumed to form a partition. C_t is the agent's credence function at the time when she starts her deliberation. Finally, U is the agent's utility or value function, which measures the utility of the elements of $\mathfrak{O}$ for the agent.[7] For each o in $\mathfrak{O}$, $U(\mathrm{o})$ is a real number. The higher the number, the higher the utility of o for the agent. Utility is assumed to be cardinal, so that for instance if $U(o_1) = 120$, $U(o_2) = 80$ and $U(o_3) = 40$, it can be concluded that o_1 is better than o_2 by the same amount as the one by which o_2 is better than o_3. There is nothing special about the value 0 or the size of the units. Evidential decision theory (and causal decision theory too) gives the same results under positive linear transformations. That is, where a is a real positive number and b a real number, multiplying each utility by a and adding b doesn't change the theory's prescriptions.

[7] I assume that the agent's utility function remains constant, so that there is no need to index it to a time.

Leaving aside some subtleties for now, EDT's recommendation in a decision situation can be expressed as follows. EDT recommends choosing any option that provides the best *evidence* or *news*, i.e. any option most likely to be accompanied by good fortune in the agent's opinion. More formally, EDT recommends to choose any option that maximizes *evidential expected utility* (EEU). In the general case of a choice between options $a_1, \ldots a_n$ with possible outcomes $o_1, \ldots o_m$, the evidential expected utility of a_i is defined as follows:

$$\mathrm{EEU}(a_i) = \sum_j C_t(o_j/a_i)U(o_j)$$

That is, $EEU(a_i)$ is an average of the values of the possible outcomes, where each such value is weighted by the agent's credence in the outcome's occurrence given a_i. EDT is usefully contrasted with Causal Decision Theory (CDT).[8] Whereas EDT's principle of rational choice makes no mention of causation, CDT's recommendations are explicitly causal. CDT recommends any option that is most likely to cause desired outcomes in the agent's opinion. More formally, CDT recommends choosing any option that maximizes *causal expected utility* (CEU). The CEU of an option a_i is defined as follows:

$$\mathrm{CEU}(a_i) = \sum_j C_t(o_j \backslash a_i)U(o_j)$$

Here $C_t(o_j \backslash a_i)$ is *not* the agent's credence in o_j given a_i. Rather, it is the agent's *causal credence* in o_j given a_i. The intent here is that $C_t(o_j \backslash a_i)$ - $C_t(o_j \backslash a_k)$ is positive just in case the agent regards o_j as more causally dependent on a_i than on a_k. The causal expected utility of a_i is an average of the values of each possible outcomes, where each such value is weighted by the agent's credence that a_i will cause a_k.

My goal is to use EDT to explain when a correlation between an act and an outcome can usefully be exploited to promote the outcomes. Now EDT is a theory of rational choice and not directly a theory of effective strategies. But as we saw in ch. 1, §4.2.4, effective strategy and practical rationality are connected. There I proposed the following connection:

[8] On CDT, see Gibbard and Harper (1978), Stalnaker (1981), Lewis (1981) and Joyce (1999), among others.

> Suppose that an agent has a choice between doing an action a or not doing a. Suppose moreover, that the agent's only goal is to have a desired outcome o occur. (It is assumed that the agent doesn't know at the time of deliberation whether o occurs.) Then if the agent knows that a is an effective strategy for o, she is rationally required to do a.

Accordingly, we can get a theory of effective strategies out of EDT by applying it to the very simple sort of decision situation just described. That is, we assume that the agent has the choice between two options, doing a certain action a and not doing a ($\sim a$). We also assume that the agent's sole goal (i.e. the only state of the world she values) is the occurrence of a certain outcome o. So the set of outcomes is $\{o, \sim o\}$, and the agent's utility function U is such that $U(o) > U(\sim o)$. For convenience we will assume that $U(\sim o) = 0$.[9] The idea, then, is to explain under which conditions a correlation between a and o is exploitable by looking at the conditions under which EDT recommends doing a in this sort of situation.

There is a complication here, however. We want to shed light on (e.g.) the exploitability of the objective correlation between throwing a rock and the window breaking by looking at what EDT says to an agent who has the choice of throwing the rock and wants the window to break. But what EDT recommends depends on the agent's subjective probabilities, not (or at least not directly) on objective probabilities. The latter will have an impact on what the agent should do only if they are somehow reflected in the agent's credence function. To ensure that objective probabilities are reflected in the agent's credence function, I will assume that the agent is a chance expert, i.e. that she knows what the chances are in our world. This gives us the required alignment between objective and subjective probabilities, for the following reason. It is a basic principle of chance that knowledge of chances constrains rational credence. For instance, if I know that the chance of the coin landing heads if tossed is $\frac{1}{2}$, then *ceteris paribus* I should have credence $\frac{1}{2}$ in the coin landing heads given that it is tossed. Lewis (1980), who was the first to explore this principle of rationality in detail, called it the *Principal Principle* (PP).

[9] Since the results of decision theory are invariant under positive linear transformations, we are always entitled to set the utility of a certain outcome at 0.

There are several different versions of the PP in the literature (see e.g. Hall (1994); Ismael (2008)). For our purposes the following one will do. Start by noting that on the conception of chance I exposed and endorsed in ch. 2, §2.2, chances are relative to reference classes. There is no such thing as the chance of y given x *simpliciter*, but only relative to reference classes for x and y respectively. This raises the following question. Suppose that an agent knows the values of $P(y/x)$ relative to various reference classes. Which of these quantities should constrain her credence in y given x? I propose the following intuitive answer, which gives us what we might call the *Relativized Principal Principle*:

> **RPP.** Let X and Y be the narrowest reference classes such that (a) the chance of y given x relative to these reference classes, $P_{Y,X}(y/x)$, is well defined and (b) at t the agent knows that x and y are of types X and Y respectively. Then if the agent is rational, her time-t credence in y given x $C_t(y/x)$ should be equal to $P_{Y,X}(y/x)$.[10]

For instance, suppose that that the agent knows that a coin is about to be tossed, that the coin is biased 60% toward heads, and nothing else. Then the agent's credence in the coin landing heads should be .6. The main reason for endorsing this proposal is its intuitive pull. Obviously if the agent knows that the coin is biased, it won't do for her to set her credence in heads to the chance of an arbitrary coin landing heads, which (let's assume) is $\frac{1}{2}$. She should set her credence to the chance of a *biased* coin landing heads. Let me briefly compare this version of the Principal Principle with Lewis's. One major difference is that mine doesn't have an admissibility clause. On Lewis's version, if an agent knows that the chance of an event y given x is p she should set her credence in y given x to p, provided that all her evidence is *admissible*. Lewis defines admissible evidence as 'the sort of information whose impact on credence about outcomes comes entirely by way of credence about the chance of those outcomes' (1980, 272). His admissibility clause is motivated by the following kind of case. Suppose that

[10] This principle can easily be extended to the case of a conditional credence involving more than two events, as follows. Let $x_1 \dots x_n$, $y_1 \dots y_m$ be events. And let $X_1 \dots X_n$, $Y_1 \dots Y_m$ be the narrowest reference classes such that (a) at t the agent knows that x_1 is of type X_1, …, x_n is of type X_n, …, Y_1 is of type Y_1, …, y_m is of type Y_m and (b) $P_{X_1,\dots,X_n,Y_1,\dots Y_n}(y_1 \dots y_m/x_1 \dots x_n)$ is well-defined. And suppose that the value of the latter quantity is p, and the agent knows this. Then if the agent is rational $C_t(y_1 \dots y_m/x_1 \dots x_n) = p$.

you know that the chance of a coin landing heads if tossed is $\frac{1}{2}$. But you also have a perfectly reliable crystal ball, which tells you that if tossed the coin will land heads. Then your credence in the coin landing heads if tossed should be 1 and not $\frac{1}{2}$. This is a case where you should not set your credence equal to the chance, because the evidence provided by the crystal ball is inadmissible. On the relativized conception of chance I proposed, we do not need such an admissibility clause. The agent knows that the toss about to take place belongs to the event-type COIN TOSS RESULTING IN HEADS. Since the chance of the coin landing heads given that it is tossed is 1 relative to this reference class, the fact that the agent should set her credence to 1 isn't an exception to my version of the Principal Principle. In this exceptional circumstance the agent should still set her credence equal to the chance, but the chance that matters is not the usual one.

To summarize, my goal here will be to explain what makes a correlation between an act a and an outcome o exploitable or not by considering what EDT says to a chance expert who has the choice of doing a and whose sole goal is to achieve o. More precisely, the goal is to explain in evidentialist terms why only those correlations that intuitively correspond to *a causing o* can be exploited. Such an explanation must be a *defense* of EDT. The reason is that EDT often *seems* to recommend acting for the sake of an outcome one cannot cause. Here is a way to see this. In the simple sort of decision situation that concerns us here, EDT says doing a for the sake of o is required just in case $C_t(o/a) > C_t(o/{\sim}a)$[11], that is, just in case the agent regards a and o as correlated. But, the thought goes, this correlation may not correspond to a causing o. The correlation between a and o could also be due to a common cause, or to a being an effect of o. In either case the correlation will intuitively not be exploitable.[12]

[11] EDT uniquely recommends doing a just in case the evidential expected utility of a is higher than the evidential expected utility of ${\sim}a$, i.e. just in case

$$C_t(o/a)U(o) + C_t({\sim}o/a)U({\sim}o) > C_t(o/{\sim}a)U(o) + C_t({\sim}o/{\sim}a)U({\sim}o) \tag{3.1}$$

Since by assumption $U({\sim}o) = 0$ and $U(o) > 0$, this simplifies to

$$C_t(o/a) > C_t(o/{\sim}a) \tag{3.2}$$

[12] Note that from the causal Markov condition, which as we have seen is a very plausible principle, it

This well-known argument against EDT is usually illustrated by a certain kind of case known as a *medical Newcomb problem*. In a medical Newcomb problem, a contemplated action is correlated with an outcome that the agent values, not because the former causes the latter but because both are independent effects of a common cause, usually some sort of genetic or physiological condition. Here is an example that we already encountered in chapter 2.

> *Chocolate.* Suppose that people who eat chocolate regularly tend to live longer. Eating chocolate, however, is causally irrelevant to life expectancy. Instead, the correlation is due to a common cause: the presence of a certain gene which causes one both to eat chocolate and (via a different causal route) to live longer.

Now imagine that Charlotte is a chance expert and is deliberating about whether to eat a Mars bar. Her sole goal is to increase her life expectancy. Since Charlotte knows that her eating chocolate correlates with increased life expectancy, she regards eating the bar as evidence for her desired outcome. Thus it seems that EDT tells her she should eat the chocolate bar, and thus to rule that the correlation between eating chocolate and increased life expectancy can be exploited to promote the latter. This is wrong, of course: since the correlation is due to the presence of a common cause, eating chocolate isn't an effective strategy for increased life expectancy. Correspondingly, in *Chocolate* Charlotte should be indifferent between eating and not eating the bar, since none of her available actions make a difference to the outcome she cares about.[13]

follows that these are the only three possibilities. In general, the causal Markov condition entails that if two events are correlated either one causes the other or they have common causes.

[13] Note that there are differences between *Chocolate* and standard examples of medical Newcomb problems, examples of which can be found in Stalnaker (1981), Eells (1982) and Price (1991) among many others. Perhaps the most well-known medical Newcomb problem is the smoking gene case, in which an agent believes that the correlation between smoking and lung cancer is due to a common genetic factor. The agent has a mild preference for smoking over non-smoking, and a strong preference against lung cancer. Clearly the agent should smoke, as doing so will have no causal influence on her getting lung cancer. But given that smoking is evidence for the gene and hence for lung cancer, EDT seems to tell the agent to refrain from smoking.

There are three differences between this case and *Chocolate*. First, in *Chocolate*, the outcome with which the action is correlated is a desirable one, not an undesirable one. Second, in *Chocolate* the agent has no intrinsic preference for eating chocolate; the only thing she cares about is increased life expectancy. Third, in standard medical Newcomb problems it is not assumed that the agent is a chance expert. These differences are irrelevant insofar as the defenses of EDT I will discuss work just as well in *Chocolate* as in standard medical Newcomb problems.

Extracting a plausible theory of exploitable correlations from EDT would amount to showing that in *Chocolate* and relevantly similar cases EDT in fact rules that spurious correlations cannot be exploited. In the next section, I will consider the most famous argument to that effect, known as the *tickle defense*. Although the defense of EDT I will provide in §4 differs from the tickle defense, it will borrow many crucial elements from it.

2 The Tickle Defense

The central idea of the tickle defense is that when the correlation between an act and an outcome is spurious, the decision-maker will necessarily have or acquire information that *screens off* the correlation, so that conditional on this piece of knowledge the act becomes evidentially irrelevant to the outcome. Thus, the tickle defense suggests a first evidentialist hypothesis as to what makes non-causal correlations unfit for acting on: they disappear when one tries to act on them. The tickle defense was originally developed by Eells (1981, 1982). As we will see Eells's original version of the tickle defense runs into problems (§1.1), but there is another version (the dynamic tickle defense) that fares better (§1.2).

2.1 Eells's Original Tickle Defense

According to Eells's original tickle defense, situations like *Chocolate* are poorly described and are in fact impossible. When I described the situation I assumed that when she starts deliberating Charlotte regards her eating chocolate as correlated to increased life expectancy. But Eells offers an argument to the effect that if Charlotte is a rational agent, when she starts deliberating she should regard what she will choose to do as independent of the gene and hence of her life expectancy.

Eells's argument relies on the assumption that any rational agent facing a decision problem must satisfy the following two conditions at the time when she starts her

Medical Newcomb problems are so-called because of their structural similarity to Nozick's (1969) Newcomb problem (see ch. 1, fn. 33). I won't discuss the Newcomb problem here, as it raises specific issues that would take me too far away from the topic of this chapter. Newcomb's problem is a very strange case anyway; given our focus on realistic cases it is harmless to ignore it.

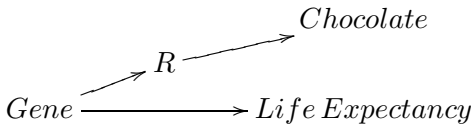

Figure 3.1

deliberation:

> **Self-Knowledge**. At the time of deliberation, the agent knows what her beliefs and desires (i.e., her credence and value functions) are. That is, $C_t(r) = 1$, where r is the agent's doxastic and evaluative state at the time of deliberation.
>
> **Control**. Which option the agent will do after deliberation is entirely causally determined by r. This entails that the agent will choose an option a just in case a appears rational in light of r. Moreover, the agent knows this.[14]

We will come back to the status of these two assumptions later on. For now let's examine their consequences in *Chocolate*. By assumption, whether Charlotte ends up eating chocolate is causally influenced by whether or not she has the gene. But if Charlotte satisfies **Control**, this means that she will end up eating chocolate just in case doing so appears rationally preferable in light of her beliefs and desires. So the gene can causally influence what she will do only by causing her to have certain credences and desires in light of which eating chocolate appears rationally preferable. (For instance, the gene might cause her to have a strong desire or 'tickle' for eating chocolate.) In other words, Charlotte's credences and desires r must be *causal intermediaries* on the causal route between the gene and her action, as in Figure 3.1.

From this and the causal Markov condition, it follows that Charlotte's credences and desires must screen off her contemplated actions from the gene. Once r is held fixed, whether or not Charlotte has the gene makes no difference to her objective probability of eating chocolate. In other words, among agents who have the same credences and desires as Charlotte, the proportion of people who end up eating chocolate is the same among those who have the gene and among those who don't. But by the **Self-Knowledge**

[14] Eells calls this second assumption 'rationality'. Since I use the word in a more ordinary sense, I have chosen a different and more suggestive name for this assumption.

assumption, when she starts her deliberation Charlotte knows her credences and desires. Moreover, since she is a chance-expert, she knows that given r, her eating chocolate and her having the gene are independent.[15] So at the start of her deliberation she should already regard her eating chocolate as evidentially irrelevant to the presence of the gene, and hence as providing no evidence either way regarding her life expectancy. If so, EDT tells her to be indifferent between eating chocolate and not doing so, as none of her actions make an evidential difference to the outcome she cares about.[16]

This argument generalizes to all cases where the correlation between an act a and an outcome o isn't due to a causing o. If a doesn't cause o, the correlation must be due either to o causing a, or to the presence of a common cause. In the former case, o must cause a by affecting the agent's beliefs and desires. Since the agent knows them, she has evidence given which the correlation disappears. In the latter case, the common cause must also cause a by affecting the agent's beliefs and desires. Again, since the agent knows them, she has evidence given which the correlation between the act and the common cause disappears. So she must regard what she will do as evidentially irrelevant to the presence of the common cause and thus to the occurrence of the desired outcome.

One problem with Eells's argument is that the assumption of **Self-Knowledge** may be too strong.[17] After all, it seems that an agent need not know her credences and desires at the start of deliberation to be in a position to deliberate. All the agent needs is to have well-defined credences and desires that can serve as inputs to her deliberation process. However, there is another version of the tickle defense that retains Eells's

[15] Eells's exact argument is more general and applies indifferently to chance-experts and to other agents, but we need not consider it here.

[16] The argument can also be put as follows. When I described *Chocolate*, I suggested that Charlotte should set her credence in her having the gene given that she eats chocolate equal to the relevant chances, relative to the reference classes HAVING THE GENE and EATING CHOCOLATE. And relative to these reference classes her eating chocolate and her having the gene are indeed correlated. But if Eells's argument is correct, the narrowest reference class for Charlotte eating chocolate that she is aware of is EATING CHOCOLATE PRODUCED BY BELIEFS AND DESIRES R. So she should set her credences to the chances relative to this reference class. And relative to this reference class her eating chocolate and her having the gene are not correlated. Eells's tickle defense thus shows that when an act and an outcome are spuriously correlated relative to certain reference classes, the agent will necessarily set her credences relative to *narrower* reference classes on which the correlation disappears. This is also true of the dynamic tickle defense to be considered in the next subsection.

[17] This objection is raised by Lewis (1981, 11). As we'll see later there are problems with **Control** too.

guiding idea but relies on a much more plausible epistemic assumption than **Self-Knowledge**. It is due in part to Eells (1984) and Horwich (1987).[18] This version relies crucially on the fact that deliberation is potentially a *dynamic* process involving several successive calculations of expected utility on the basis of new information provided by earlier calculations. On this 'dynamic' version of the tickle defense, although Charlotte regards her eating chocolate and her having the gene as correlated at the start of her deliberation, she will during her deliberation acquire information given which the correlation disappears. By contrast to Eells's original version, the piece of knowledge that ensures independence isn't present at the outset. Rather, it is gained during the very process of deliberation.

2.2 The Dynamic Tickle Defense

To introduce the dynamic tickle defense, let's imagine that the gene causes chocolate-eating by causing its bearers to have certain inclinations or beliefs in light of which eating chocolate appears obviously preferable after quick deliberation. Agents who have the gene tend to judge, after a single calculation of expected utility, that eating chocolate maximizes their expected utility. And a significant proportion of those agents act on this judgment, so that on the whole the gene-bearers are more likely to end up eating chocolate. If the gene works solely by influencing the results of initial deliberation, those results should screen off chocolate-eating from the gene. That is, among people who initially judge that eating chocolate is preferable, those who go on to act on this judgment are no more or less likely to have the gene than those who end up not eating chocolate. Likewise, among people who initially judge that eating chocolate is *not* preferable, those who go on to act on this judgment are no more or less likely to have the gene than those who end up not eating chocolate. By hypothesis Charlotte knows this, since she is a chance expert.

Now imagine Charlotte starting her deliberation. She calculates her expected utilities for eating chocolate and not doing so on the basis of her utility function U and

[18] The version I will present doesn't correspond exactly to Eells's (1984) and Horwich's arguments, however.

her credence function C_t, and arrives at a judgment j_t of the form 'The EEU of eating chocolate is x and the EEU of not eating chocolate is y'. For definiteness, let's assume that j_t is a judgment that favors eating chocolate. (The argument works equally well if j_t favors not eating chocolate, or favors both options equally.) Does EDT tell Charlotte to act on j_t and eat chocolate? Plausibly not, for the following reason. EDT would recommend acting on j_t if it were committed to the following principle:

> **Current Opinion Fixes Action**.[19] If C_t characterizes the agent's credence function at time t, then at t the agent is rationally required to perform an act a just in case a maximizes her evidential expected utility calculated on the basis of C_t (her time-t evidential expected utility, for short).

But EDT charitably understood isn't committed to this principle. In fact, no plausible decision theory is committed to the principle that an agent should do an act a just in case a maximizes her present expected utility (be it causal or evidential). Joyce (2012) shows this by considering the following case, in which both EDT and CDT give the same results[20]:

> Imagine a Blackjack player who has seen her top card (a seven) and the dealer's top card (an eight), but who has yet to peek at her hole card, which she can do cost-free. The player knows that she should stand pat if her cards total 17 or more, and that she should ask to be 'hit' with another card if they total 16 or fewer. Suppose she calculates her chance of having at least 17 without looking at her hole card, and finds it to be 0.4, so that the [evidential or causal] expected payoff of taking a hit exceeds that of standing pat. While this is fine as an academic exercise, if the player took a hit on this basis we would think her daft. Even though she can assess probabilities and utilities without factoring in the hole card, she clearly should not act on such assessments. (2012, 126)

Rather, since the costs of looking at her hole card are negligible, she should first take a look at the hole card so as to gather all the easily available relevant evidence before making a decision. The case shows that no plausible decision theory can be committed to the principle that current expected utility calculation fixes what the agent should do. Current utility judgments determine what the agent should do only when those judgments incorporate all the easily available evidence relevant to what the agent should do.

[19] I borrow the phrase from Joyce (2012).

[20] Joyce is concerned with showing that CDT isn't committed to **Current Opinion Fixes Action**, but his argument easily extends to EDT. Joyce uses this principle to show that CDT correctly handles Egan's (2007) alleged counterexamples to it.

Coming back to Charlotte, we can now see that if she were to act on j_t, she would violate this principle, as j_t doesn't incorporate all the easily available evidence. *In particular j_t doesn't incorporate the evidence that it itself provides.* j_t is evidence relevant to Charlotte's decision because it provides information about the presence of the gene that *screens off* the correlation between the gene and eating chocolate. Given that she initially prefers chocolate, Charlotte is likely to have the gene, whether or not she ends up eating chocolate later. Moreover, j_t is easily available evidence: Charlotte need only pay minimal attention to her own deliberation to acquire this crucial piece of information. So it would be foolish for Charlotte to act on j_t and eat chocolate. Were she to do so, she would thereby ignore easily available evidence in light of which she has no reason anymore to eat chocolate.

These considerations show that we need to modify our earlier formulation of EDT's recommendations. In section §1 I said that EDT recommends choosing the act that maximizes evidential expected utility relative to the agent's credence function *at the time of deliberation*. The problem with this formulation is that the agent's credence function might not remain the same through deliberation. The reason is that the agent's very deliberation might produce new information which the agent should incorporate in her all-things-considered assessment of what she should do. This means that the agent's deliberation might have to go through several steps before EDT issues a recommendation to act in a certain way. This dynamic deliberation process can be modeled as follows.[21] The deliberation is composed of a series of stages $\{0, 1, \dots\}$. Each stage i comprises two steps. During the first step, the agent assesses her evidential expected utilities for her options in light of her time-i credence function C_i. She thus arrives at a judgment j_i of the form '$\mathrm{EEU}_i(a) = x$ and $\mathrm{EEU}_i(\sim a) = y$'. During the second step, the agent updates on j_i so that her new credence function C_{i+1} is equal to $C_i(\bullet/j_i)$. Once this second step is completed, the agent returns to the first step. The process stops when the agent arrives at an equilibrium assessment of her expected utilities - that is, when she reaches a stage n such that $\mathrm{EEU}_n(a) = \mathrm{EEU}_{n+1}(a)$ and $\mathrm{EEU}_n(\sim a) =$

[21] This model is inspired by the one built by Skyrms (1990) for dynamic causal decision theory. Joyce (2012) offers a very clear presentation of Skyrms's model.

$\mathrm{EEU}_{n+1}(\sim a)$. At this point all the evidence that the deliberation might produce is in, and EDT issues a recommendation: it says to do any action that maximizes evidential expected utility according to the agent's equilibrium assessment.

In Charlotte's case, this deliberation process pans out in the following way. Once Charlotte has learned her initial assessment j_t, she should reassess her expected utilities on the basis of her new credence function $C_{t+1} = C_t(\bullet/j_t)$. Since given j_t her eating chocolate provides no evidence anymore for the presence of the gene, she will reach a new judgment j_{t+1} to the effect that the evidential expected utilities of eating and not eating chocolate are equal. Moreover, this judgment is an equilibrium assessment, since given j_{t+1} eating chocolate is still evidentially irrelevant to the gene. Consequently, EDT correctly tells her to act on this judgment, and thus to be indifferent between her two options. The outcome is the same as in Eells's original tickle defense. But the 'tickle' that screens off the spurious correlation between eating chocolate and higher life expectancy isn't Charlotte's original credences and desires, but her initial assessment of her expected utilities.

It is useful to compare the assumptions of this dynamic tickle defense with the assumptions of Eells's original tickle defense. The dynamic tickle defense doesn't presuppose **Self-Knowledge**, and in that respect escapes the objection against Eells's original defense. Eells's controversial requirement has been replaced with a much more plausible epistemic requirement: the agent should take into account all easily accessible evidence (including the results of the successive stages of her deliberation) when assessing what she should do. As the case of the blackjack player shows, this requirement is very plausible. The dynamic tickle defense contains an implicit assumption similar to **Control**. Above we implicitly assumed that the gene causes eating chocolate solely by affecting the results of deliberation (more precisely, by causing its bearers to judge after initial deliberation that eating chocolate is preferable). This ensures that by learning the results of her deliberation, the agent thereby acquires evidence that decorrelates her eating chocolate from the gene. If the gene were to affect what agents do via a route that bypasses deliberation, deliberation could not act as a screen between the act and the outcome. (This will be important later on in the chapter.)

At the start of this subsection we assumed that the gene affects deliberation in a simple way, solely by affecting the results of agents' *initial* deliberations. It should be easy to see that the dynamic tickle defense generalizes to more complex (and more fanciful) ways in which the gene might influence deliberation. Suppose for instance that the gene influences deliberation not only by making the agents more likely to initially judge that eating chocolate is preferable; in addition, the gene causes those deliberators who engage in a second round of expected utility calculations to overestimate the utility of eating chocolate. In that case the gene-chocolate correlation doesn't entirely disappear given the results of initial deliberation. Among those people who initially regard eating chocolate is preferable, those who go on to recalculate their utilities are still a bit more likely to end up judging again that eating chocolate is preferable if they have the gene than if they don't. But this simply delays the moment at which Charlotte will end up regarding eating chocolate as evidentially irrelevant to the gene. Once she arrives at her initial judgment j_t and updates on it, Charlotte still regards eating chocolate as somewhat evidentially relevant to the presence of the gene, although less than at the beginning of deliberation. So when she assesses her time-t+1 utilities, she will still arrive at a judgment (j_{t+1} to the effect that eating chocolate is preferable. But once she updates on *that* judgment, she now regards her contemplated action as providing no additional evidence for the presence of the gene, for together j_t and j_{t+1} screen off the correlation. More generally, as long as the gene influences what Charlotte will do solely by influencing the results of her deliberation up to a finite stage t+n, when she reaches that stage Charlotte will have acquired evidence that entirely decorrelates the gene from eating chocolate. But what if the gene continuously influences deliberation, so that there is no stage at which its influence stops? Then at no point during deliberation will Charlotte acquire evidence that completely decorrelates her contemplated act from the gene. It seems to me, however, that such cases are sufficiently fanciful and unrealistic to be properly ignored in the context of formulating a theory concerned with explaining effective strategies in the actual world. I will therefore ignore them in what follows.

3 Price's Evicausalism

The tickle defense gives us a first hypothesis as to why a correlation between an act and an outcome cannot be exploited only when the outcome doesn't causally depend on the act. In such cases, the correlation is bound to *disappear* when we try to act on it, in the sense that deliberation necessarily provides information that screens off the correlation. Perhaps, then, the tickle defense can be put to work to give us what we want: a satisfactory evidential explanation of what makes a correlation exploitable or not, on the basis of which we could extract a solution to Russell's problem. This idea is developed by Price (1996; 2007; 2012; 2014), partly in collaboration with Brad Weslake (Price and Weslake, 2009). Price's motivations are very similar to mine. One of his main claims is that a solution to Russell's problem must not make the practical relevance of causation a mystery, and that EDT is therefore a good starting point since its principle of rational choice has an air of self-evidence.[22] Price's theory of what makes causal correlations exploitable is most clearly expressed in his (2012). There Price simply *identifies* causal dependence with the sort of probabilistic dependence that is good for a chance expert to act on according to the tickle defense - i.e., those correlations that disappear in the context of deliberation. Price calls this view 'Evicausalism'. It can be expressed a bit more precisely as follows:

> **EC**. e causally depends on c just in case a chance expert who is deliberating about c and who wants e to occur would regard c as evidence for e (at least at the end of her deliberation).

For instance, on **EC** what makes the window breaking causally dependent on throwing a rock at it is that a chance expert deliberating about the latter and who wants the former to occur would regard throwing the rock as evidence for the window breaking at the end of deliberation. To defend the idea that **EC** will count all and only those correlations that we intuitively think as good for acting on (and thus will be an extensionally adequate theory of effective strategies), Price appeals to the tickle defense. Thus he writes that 'the acknowledged successes of the Tickle Defense do much to meet the

[22] Remember Papineau's quote cited in chapter 2: 'Doesn't everybody want it to be probable that they will get what they want?' (2001a, 244).

objection that there are cases in which it is obvious that [EC] will attribute causal dependency, where actually there is none' (2012, 513). As Price recognizes, **EC** is only a first pass at a proper account of causal dependence.[23] The reason is that many causal dependence relations involve events that we cannot manipulate (and thus cannot deliberate about), so that these relations are not captured by **EC**. So **EC** needs to be extended to unmanipulable causes. He says little about to do so, but his idea seems to be that we should endorse the interventionist strategy described at the beginning of this chapter. That is, we can identify causal correlations in general with those correlations that survive under a third event (a sort of intervention) that is relevantly like a deliberation in that it performs the same task of screening off spurious correlations.

It seems to me that there are two problems with Price's view. Both problems are tied to the fact that the tickle defense relies on a crucial *causal* assumption. This is the assumption that what a deliberating agent will do is entirely *causally determined* by the results of her deliberation. It is present both in Eells's original tickle defense (through **Control**) and in the dynamic version. Let's call it the *assumption of control*. As we have seen, it is essential to ensure that in the case of a spurious correlation between an act and an outcome, the agent's deliberation will provide information that screens off the correlation. In cases where the agent's action is not entirely causally determined by her own deliberation, non-exploitable correlations may not be screened off by her deliberation. To illustrate, let us consider a hypothetical scenario presented by Papineau (2001a, 253-4) as part of an argument against EDT[24]:

> *Smoking.* Suppose that whether or not one smokes is a function of two factors: whether one judges after deliberation that smoking is in one's best interest, and the probabilistically independent presence of a psychologically undetectable chemical in the bloodstream. Even if you judge that smoking isn't the right thing to do, you might still find yourself unwittingly reaching for the cigarette if you have the chemical. For definiteness, we may imagine that one is 99% likely to smoke if one's deliberation favors it and if one has the chemical; 90% likely if one's

[23] See especially Price and Weslake (2009, §6.3).

[24] I have made some modifications to the case. Lewis (1981, 11) mentions in passing the possibility of such cases as an objection to the tickle defense, but as far as I know Papineau was the first to discuss in detail cases where the assumption of control isn't satisfied.

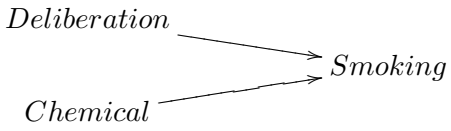

Figure 3.2

> deliberation favors smoking and one doesn't have the chemical; 10% likely if one's deliberation doesn't favor smoking but one has the chemical; and 1% likely if deliberation doesn't favor smoking and the chemical is not present.

This case (which should not be too hard to imagine if you have addictive tendencies) has the causal structure represented in Figure 3.2. Now imagine a chance expert (Samantha, say) who is deliberating about whether to smoke and whose sole goal, for whatever reason, is to have the chemical. Here the correlation between the cause and the effect doesn't disappear during deliberation, since the chemical causes smoking by a route that bypasses deliberation. Consequently, here the tickle defense doesn't work. However her deliberation goes, Samantha will at any point of her deliberation regard smoking as positive evidence for the presence of the chemical. This shows how crucial the aforementioned causal assumption is to the tickle defense's success.

The fact that the tickle defense relies crucially on the assumption of control creates the following problem for Price's account. Given this assumption, it is not clear that **EC** provides a satisfactory explanation of what makes correlations exploitable or not. To see why, note first that since the chemical-smoking correlation doesn't disappear during deliberation, **EC** runs into the danger of mistakenly counting the correlation as exploitable in *Smoking*. As far as I can see, there are only two options available to Price for avoiding this unwelcome result. The first is to posit that we can only in fact deliberate about options that are under our full control, as the tickle defense assumes.[25] If so, in *Smoking* Samantha cannot deliberate about whether or not to smoke in the

[25] In fact, this assumption is not peculiar to proponents of the tickle defense. Causal decision theorists also often assume that to count as an option for deliberation an action must be under the agent's full control. For instance, in 'Causal Decision Theory' Lewis writes: 'Suppose we have a partition of propositions that distinguish worlds where the agent acts differently... Further, he can act at will so as to make any one of these propositions hold; but he cannot act at will to make any proposition hold that implies but is not implied by (is properly included in) a proposition in the partition. The partition gives the most detailed specifications of his present action over which he has control. Then this is a partition of the agents' alternative *options*' (1981, 7).

first place, so that **EC** doesn't count smoking as an effective strategy for influencing the presence of the chemical. The second is to assume that one *can* deliberate about whether or not to smoke in *Smoking*, but only under the fictitious assumption that whether or not one smokes is under one's complete control. (So one assumes during deliberation that the chemical has no influence over what one will do.) If so, any agent deliberating about smoking must regard her smoking as providing no evidence for the chemical's presence, so that **EC** doesn't count smoking as a cause of the chemical after all. Both options are problematic, however. Regarding the first one: it is unclear that we ever have complete control over any of our actions. After all, complete control presumably requires nomological determination. But as we have seen in ch. 1, only global states of the world can nomologically determine anything. (There is always the possibility of an intervention from afar that prevents us from accomplishing the action that is the object of deliberation.) So the first option implausibly entails that we cannot deliberate about anything. The second option doesn't run into this problem: even if we do not ever have complete control over what we will do, we may still be able to reason as if we did. But the problem here is that the requirement that an agent should deliberate as if she had complete control over her options is mysterious. Why should we reason under the fiction of complete control in the context of deliberation? Price claims that since **EC** accounts for the exploitability of causal correlations in terms of the self-evident evidential principle of rational choice, 'it easily explains the *practical* relevance' of causal facts (2012, 485). But if Price must also take on board the mysterious assumption that one should deliberate under the fiction of complete control, his explanation turns out to be far less clear than advertised.

A second problem tied to the assumption of control arises when we turn to the task of extending **EC** into an account that handles unmanipulable causes as well. Price's suggestion, remember, is that we can identify causal correlations with those correlations that survive under an event (a kind of intervention) that is *relevantly like* deliberation, and in particular performs the same job of screening-off spurious correlations. The problem here is that for deliberation to screen off spurious correlations, it is crucial that what the agent will do be causally determined by her deliberation. This suggests

that, similarly, to count as a Pricean intervention an event will have to satisfy certain explicitly causal characteristics. In particular, a Pricean intervention on an event a will have to causally determine a's occurrence, just like Pearl's and Woodward's interventions do. Unless Price gives us a way to capture the idea of causal determination in non-causal terms, the resulting theory will not be reductive and thus cannot work as a solution to Russell's problem, contrary to what Price claims.[26]

4 A New Defense of EDT

The considerations of the previous section show that in order to articulate an evidentialist theory of exploitable correlations that can serve my (and Price's) purposes, we need to provide a defense of EDT that doesn't rely on the assumption of control. This is the task of the current section. I will start by providing an evidentialist explanation of why the smoking-chemical correlation in *Smoking* cannot be exploited (§4.1), and then generalize it to all cases of spurious correlations (§4.2).

Note that in *Smoking*, although the agent's deliberation doesn't causally *determine* what she will do, it still has a causal influence on whether the agent will smoke. As we will see, this is important to explain why the smoking-chemical correlation cannot be exploited. More generally, in this section, I will provisionally assume that for an act a to be a proper object of deliberation, the agent's deliberation must be a cause (if only a partial one) of whether she will do a. We might call this the *assumption of option causation*. But as we have seen, for our purposes it is important that at the end of the day the characteristics of deliberation that explain why certain correlations are good for acting on and others not are cashed out in non-causal terms. n §5, I will show that the assumption of option causation can be replaced with a non-causal description of the relations between deliberation and options.

Although my explanation doesn't presuppose that an agent's options must be under her full control, it nevertheless relies on a causal assumption: the assumption that for an action a to be a proper object of deliberation, the agent's deliberation must be a

[26] See e.g. Price and Weslake (2009, §6.3).

cause (if only a partial one) of whether she will do *a*. (This assumption is satisfied in *Smoking*, since the agent's deliberation has *some* causal influence on whether she will smoke.) We might call this the *assumption of option causation*. But as we have seen, for our purposes it is important that at the end of the day the relations between an agent's deliberation and her options be characterized in non-causal terms. In §5, I will show that the assumption of option causation can be replaced with a non-causal description of the relations that must hold between deliberation and options for choice.

4.1 *Smoking* Revisited

Consider *Smoking* again. The goal is to explain in evidentialist terms why the correlation between smoking and the chemical cannot be exploited, despite the fact that it doesn't disappear during deliberation. My hypothesis is that this is because the presence of the chemical is *probabilistically independent* of the results of the agent's deliberation. That is, whether or not the agent judges to smoking to be preferable isn't correlated with her having the chemical. We might also put the relevant point like this. In *Smoking*, the agent judging that smoking is preferable is positively correlated with her smoking. If the agent judges that smoking is in her best interest, she is more likely to smoke. *But if she judges that smoking is best, she is not more likely to have the chemical.* The correlation between deliberation and smoking doesn't 'translate' into a correlation between smoking and the chemical. This probabilistic independence was stipulated in the vignette with which I introduced *Smoking* in §3. We will shortly explore the consequences of relaxing it. First I will explain how this probabilistic independence makes the correlation between smoking and the chemical unexploitable.

Instead of considering directly the question whether Samantha should smoke, consider the following question: should Samantha *deliberate* about whether to smoke? Note that deliberation is a form of intentional action, and so the sort of thing which it can be rational to do or not. The consequence of the aforementioned probabilistic independence is that, according to EDT, Samantha is not *rationally required* to deliberate about smoking. The reason is as follows. By hypothesis, the outcomes of Samantha's

deliberation are uncorrelated with the outcome she desires to obtain (the presence of the chemical). Whatever judgment she will arrive at at the end of her deliberation, she is no more or less likely to have the chemical. This means that Samantha's deliberation cannot provide her with any positive evidence regarding the outcome she cares about. Since EDT says that an action is rationally required just in case it provides evidence for good results, by its own lights Samantha is not rationally required to deliberate about whether to smoke. She might just as well toss a coin and make up her mind about what to do on the basis of the toss's outcome. But if this is so, EDT cannot also say that smoking is rationally required because of its correlation with the chemical. For suppose that EDT *does* issue this recommendation. Then on the one hand EDT says that one of the two possible actions (smoking and not-smoking) are rationally required. On the other hand, it says that it is pointless for EDT to deliberate about those actions, and that she might as well use a non-deliberative decision procedure on which she is equally likely to end up smoking and to end up not smoking. There is a strong tension between these two claims. Now is not impossible for a decision theory to both recommend an act and to also advise against deliberating about the act, for instance if the deliberation is too complex or too time-consuming for the agent. But in those cases this will be because deliberating has certain costs for the agent - it is associated with outcomes she values negatively. But here by hypothesis the only outcome Samantha cares about is whether she has the chemical. So we cannot explain away the tension between EDT's two claims by appealing to costly features of deliberation. The tension remains, and it is sufficiently strong to make the following principle plausible: if EDT recommends an action on the basis of its correlation with a desired outcome, it cannot also tell the agent that it is pointless to deliberate about whether to do the action (at least if the relevant outcome is the only one that the agent ultimately cares about). If so, given that EDT tells Samantha that deliberating is pointless, it cannot also say that the smoking-chemical correlation is a good reason to smoke.

This explanation of why EDT doesn't tell Samantha to smoke in *Smoking* naturally invites the following answer. What if we change the details of *Smoking* to make it so

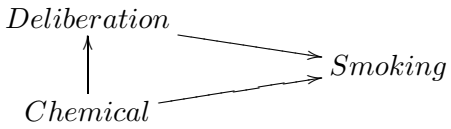

Figure 3.3

that the presence of the chemical is correlated with the results of Samantha's deliberation? For instance, let us now stipulate that Samantha is more likely to have the chemical if she judges that smoking is preferable. Then since the results of Samantha's deliberation can now provide her with evidence regarding the presence of the chemical, the previous argument doesn't apply anymore. Since deliberating about smoking can provide evidence that Samantha has the chemical, EDT doesn't entail anymore that it is pointless to deliberate. However, as we will now see, the explanation I proposed in fact generalizes to cases where deliberation and chemical are correlated.

The argument goes as follows. Note first that for there to be a correlation between the chemical and the results of deliberation, there would have to be a *direct causal connection* between the two. As we saw when discussing SGS in ch.2, causes of a joint effect are not correlated with each other, unless one causes the other or they have a common cause. So for there to be a correlation, it would have to be either that (a) the chemical causally influences the results of deliberation or (b) there is a common cause of both or (c) the results of deliberation causally influence the presence of the chemical. But in each of these cases, we can apply the tickle defense to show that during deliberation, Samantha will acquire evidence that screens off the presence of the chemical from the remainder of her deliberation. After having acquired this evidence, Samantha will regard her future assessments of the utilities of her acts as uncorrelated with the presence of the chemical, so that we can then apply the argument I just gave to show that EDT doesn't recommend smoking after all.

Let's see how this goes by considering case (a) first. Its causal structure is represented in Figure 3.3. Let's assume for now that the chemical influences how deliberation goes solely by influencing the *initial* results of deliberation. Perhaps the chemical causes agents to have a strong desire to smoke, in light of which after initial calculation of expected utilities smoking appears preferable. Now imagine that Samantha starts

her deliberation about smoking, and arrives at the initial judgment j_t that smoking is preferable.[27] Note first that EDT doesn't recommend Samantha to smoke on the basis of j_t, because j_t isn't an equilibrium assessment of the relative preferabilities of her options. The reason is that once she learns j_t, Samantha should regard smoking as less evidentially relevant than before. Before learning j_t, Samantha regarded smoking as evidence for the chemical for two reasons: because the chemical directly causes smoking, and because her smoking is evidence that she initially judged that smoking is preferable, which in turn is evidence that she has the chemical. But since she now knows that she initially judges that smoking is preferable, the second evidential route from smoking to the chemical is screened-off: after having learned j_t, Samantha regards smoking as evidence for the chemical only because the chemical directly causes smoking, not because it affects deliberation. Consequently, j_t constitutes information that changes the evidential relevance of smoking to the chemical. So EDT tells Samantha that she should recalculate the utilities of her options. But note that upon learning j_t, Samantha should now regard the remainder of the deliberation as evidentially independent of the presence of the chemical. Since the chemical can affect her deliberation solely by influencing her initial judgment, once she has updated on the latter Samantha knows that she is equally likely to have the chemical, whatever she ends up judging during the remainder of her deliberation. In particular, she is no more likely to have the chemical if she ends up judging that smoking is preferable than if she ends up judging that both options are on a par. At this point, the argument I proposed in the original version of *Smoking* applies again: EDT now tells her that it is pointless to continue her deliberation about whether to smoke. Thus, for the reasons pointed out above, EDT cannot properly recommend Samantha to smoke in this case.

Three remarks about this argument. The first is that it makes use of the tickle defense, but for non-standard purposes. Here the tickle defense isn't used to show that the agent will during deliberation acquire information that screens off the correlation between her contemplated actions and their causes (here the chemical). Rather, it

[27] The reasoning works equally well if instead she arrives at the judgment that not-smoking is preferable, or that both options are on a par.

is used to show that at some point during her deliberation, the agent will acquire information that decorrelates the chemical from *the remainder of her deliberation*.

The second remark is that this argument generalizes to cases where the chemical affects deliberation in more complex ways than by solely affecting initial judgments.[28] Suppose for instance that the chemical affects deliberation not only by influencing initial judgments, but by making agents more likely to overestimate the utility of smoking if they engage in a second utility calculation. Then although Samantha will still regard the chemical as correlated with the remainder of her deliberation upon learning her initial judgment, the correlation will be screened off once she learns her second utility judgment.

The third remark is that this argument also generalizes to cases (b) and (c). Suppose first that the correlation between the chemical and deliberation is due to a common cause c, as in Figure 3.4. By the same reasoning, Samantha should during deliberation acquire information that decorrelates the remainder of her deliberation from the common cause. But since the correlation between the chemical and deliberation is entirely due to the common cause, she will thereby come to regard the presence of the chemical as independent of her future judgments regarding the utilities of her options.

Consider case (c) next, whose causal structure is represented in Figure 3.5. Suppose for simplicity that the causal influence of deliberation on the chemical goes solely through agents' initial assessments of their preferences. When an agent initially judges that smoking is preferable, this causes the chemical to be released in her bloodstream; later reassessments of her preferences have no causal influence on the presence of the chemical. Then by learning her initial assessment, Samantha gains information given which the evidential relevance of smoking to the chemical is decreased: she now regards smoking as evidence for the chemical only because the chemical directly causes smoking. The second evidential route from smoking to the chemical that goes through deliberation is now 'blocked'. Consequently, Samantha's initial judgment isn't an equilibrium assessment, and EDT tells her to recalculate. Moreover, now that she knows her initial judgment Samantha regards the remainder of her deliberation as evidentially

[28] See the end of §2.2 above.

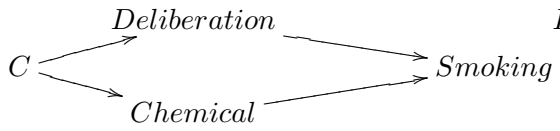

Figure 3.4

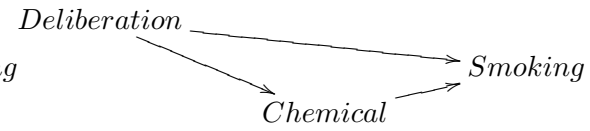

Figure 3.5

independent of the presence of the chemical. Here again, by the same reasoning as above, EDT cannot properly tell her that she should smoke. This reasoning generalizes to more complicated hypotheses about the causal effects of deliberation on smoking. Suppose for instance that Samantha's second calculation of expected utility also has a causal impact on the presence of the chemical. Then her initial judgment won't decorrelate the remainder of her deliberation from the presence of the chemical. But after her second calculation, Samantha will then have evidence given which the remainder of her deliberation is again screened off from the presence of the chemical. More generally, as long as the influence of Samantha's judgments on the presence of the chemical *stops* at some point during the deliberation, then when Samantha reaches this point she will have evidence given which the remainder of her deliberation and the presence of the chemical are independent. What if, however, the causal influence of deliberation on the chemical never stops? What if, for instance, every judgment that Samantha might potentially reach during her deliberation in some way causally influences the presence of the chemical? Then at any point in her deliberation Samantha will still regard the remainder of her deliberation as correlated with the presence of the chemical. Such cases, however, seem to me sufficiently fanciful and unrealistic to be ignored in the context of a theory concerned with explaining effective strategies in the actual world. In what follows, I will therefore ignore such cases of continuous influence.

I conclude that my explanation of why the smoking-chemical correlation cannot be exploited in the original version of *Smoking* generalizes to cases where the results of deliberation and the presence of the chemical are correlated. In those cases, the correlation isn't *robust* under deliberation: it is bound to be screened-off by the evidence provided by Samantha's own utility judgments. Samantha may start in a state where

she regards her deliberation as potentially providing positive evidence for the presence of the chemical. She is guaranteed to end up in a state where the further results of her deliberation are guaranteed to be evidentially irrelevant to the presence of the chemical. At this point EDT tells her that it is pointless for her to deliberate any further about whether to smoke, so that it cannot claim at the same time that the smoking-chemical correlation is a good reason to smoke. Our survey of cases (a), (b) and (c) also makes the following claim plausible. Suppose that a chance expert is deliberating about whether to do a for the sake of an outcome o. Then during her deliberation the agent will acquire evidence that screens off any correlation between her deliberation and o that comes from a direct causal connection between them - i.e., a causal connection that doesn't go via the deliberation's causal influence on a. This will be important in the next subsection.

4.2 A General Evidentialist Explanation of Exploitable Correlations

We are now in a position to propose a general evidentialist explanation of the distinction between exploitable and unexploitable correlations that, by contrast to the tickle defense, doesn't presuppose any assumption of control. The idea is that what makes a correlation unexploitable isn't that it disappears during deliberation - this isn't true in *Smoking*, for instance. Rather, the mark of an unexploitable correlation between an act a and an outcome o is that o is uncorrelated with the results of deliberation about a (or that the correlation between deliberation and outcome disappears during the deliberation itself). The explanation I propose for why this makes the correlation unexploitable is as follows. If the results of deliberating about a are uncorrelated with the agent's desired outcome, EDT tells the agent that it is pointless for her to deliberate about a; she might as well flip a coin. Consequently, EDT cannot also tell the agent that the correlation between a and o is a good reason to do a.

This evidentialist explanation applies in cases where the assumption of control isn't satisfied and where consequently the tickle defense is silent. But note that it *also* applies in cases for which the tickle defense was designed. Consider *Migraine* again, assuming for simplicity that the gene influences deliberation solely by influencing its

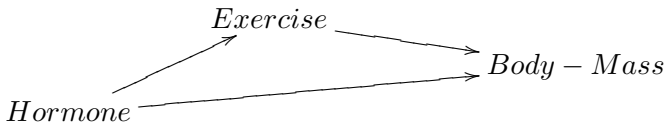

Figure 3.6

initial results. Upon learning her initial assessment of expected utilities, Charlotte comes to regard eating chocolate and the gene as independent. But she also come to regard the remainder of her *deliberation* as independent of the gene. That is, she now knows that her future deliberation judgments cannot provide her with any evidence regarding the presence of the gene. So here my explanation applies: EDT tells her that it is pointless to deliberate any further about whether or not to eat chocolate. Thus, my evidentialist explanation of what makes a correlation exploitable or not is preferable to the tickle defense as it handles both the cases for which the tickle defense was designed and those in which it is silent.

There is one addendum we need to make to this evidentialist theory of exploitable correlations. So far I have talked as if whether a correlation is exploitable is an all-or-nothing matter. This isn't the case, as the following example illustrates. Suppose that the correlation between exercising and increased body mass is due in part to the fact that the former causes the latter, and in part to a certain hormone that causes one to exercise a lot and (by a different route) to have increased body mass. The causal structure is represented in Figure 3.6. There are, so to say, two components of the correlation between exercising and body mass. One is due to the causal influence of exercising on body mass and can therefore be exploited. The other is due to the common cause and is therefore not exploitable. So we need to extend the present explanation of exploitable correlations into a theory of which *parts* of a correlation can be exploited. The natural way to do this is as follows. Consider a chance expert deliberating about whether to exercise, and whose sole goal is to increase her body mass. Since both the deliberation and the hormone are causes of exercising, the results of deliberation and the hormone should be independent of each other.[29] That is, during deliberation

[29]If they are correlated, this will be due to a causal connection between deliberation and hormone, which will be screened-off during deliberation.

the agent knows that whatever she ends up judging at the end, she is no more or less likely to have the hormone. This means that on EDT it is pointless to deliberate about whether to smoke so as to influence one's body mass by way of influencing the presence of the hormone. Correspondingly, EDT cannot also say that the fact that exercising is correlated with the hormone (and thus with the desired outcome) is a good reason to exercise. By EDT's own lights, then, the presence of the hormone isn't the sort of thing that the agent can usefully influence (at least not in the present context), and so it is the sort of thing that she should hold fixed when assessing how much exercising would advance her desired goal. Correspondingly, the presence or absence of the hormone can be held fixed when estimating how much the correlation between exercising and increased body mass can be exploited.

On the basis of this evidentialist theory of exploitable correlations, we can now start explaining why an act a is a good way to influence an outcome o when a causes o, but not when a is an effect of o. Remember that we are provisionally assuming that for a to be a proper object of deliberation, deliberation must be causally relevant to the action. Given this constraint, when an outcome causes an act there are two main possibilities. Either the outcome causes the act by influencing the deliberation, as in Figure 3.7. Or it causes the act via a route that bypasses deliberation, as in Figure 3.8. In the former case, as we have just seen, the agent will acquire evidence that screens off the cause both from her contemplated action and from the remainder of her deliberation, so that EDT tells her it is pointless to deliberate. In the latter case, deliberation and outcome will be uncorrelated from the start, unless there is a causal connection between them. But as we have seen in §4.1, this causal connection will itself be screened off by evidence acquired during the deliberation process. Either way the agent will come to her deliberation as uncorrelated with the desired outcome, so that EDT cannot tell her that the act-outcome correlation is a good reason to do the act. This explanation generalizes to cases where the correlation between an act a and an outcome o is *spurious*, i.e. due to a common cause. By the same reasoning, EDT won't recommend doing the act so as to influence the outcome by influencing the occurrence of the common cause.

Figure 3.7

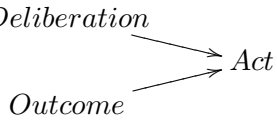

Figure 3.8

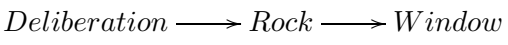

Figure 3.9

Now, contrast this with a case where the act is a cause of the outcome. Suppose for instance that Suzy is deliberating about whether to throw the rock in order to break the window. The causal structure is represented in Fig. 3.9. There the correlation between act and outcome won't disappear during deliberation. At each stage of her deliberation, Suzy regards throwing the rock as evidence for the window breaking. Moreover, she regards her *deliberation* as itself correlated with the outcome. That is, at any stage of deliberation, she believes that if she ends up her deliberation with the judgment that throwing the rock is best, the window is more likely to break (because the judgment is evidence that she will throw the rock, and thus that the window will break). Consequently, EDT will tell her that the correlation between throwing the rock and the window breaking is a good reason to throw the rock. Here the crucial factor that distinguishes this case from *Smoking* is the fact that Suzy's judgments about what she should do are correlated with the occurrence of the desired outcome, and remain so even given the evidence she might acquire during deliberation. This suggests, then, that the fundamental asymmetry between cause and effect that explains the asymmetry of effective strategies is the following one. When an act a causes o, an agent's deliberation about a will be correlated with o, and robustly so. But when an act a is an effect of o, the agent's deliberation will not be correlated with o, or at least not robustly so. That is, any such correlation will disappear during deliberation. The suggestion I will develop in the next chapter is that this asymmetry is constitutive of the difference between cause and effect.

5 Deliberation and Options

In the previous section, I offered a general evidentialist explanation of what makes a correlation exploitable or not. I argued that this explanation entails that a correlation between an act and an outcome is exploitable only when the act is intuitively a cause of the outcome. I also started identifying the features of deliberation that make only causal relations exploitable. The crucial feature is that when the outcome is not an effect of the action, a chance expert will regard her deliberation about the act as uncorrelated with the occurrence of the outcome. However, my explanation for this latter claim hinged on the assumption that when a is an option, deliberation about a must be a cause of whether the agent will do a. This is what I called the assumption of option causation. Now if my sole purpose here were to provide a theory of exploitable correlations, this causal assumption would be unproblematic. A theory of exploitable correlations need not be free of causal assumptions, as long as those assumptions are reasonable. This is what distinguishes the assumption of option causation from the assumption that when we deliberate what we will do is *entirely* under our causal control. As we have seen, the latter assumption is problematic since we presumably do not have complete control over anything. By contrast it is obvious that our deliberations typically are causally relevant to what we will do. But my purpose here is not to give a plausible theory of exploitable correlations only. The strategy I proposed at the beginning of this chapter is to identify causal correlations *in general* with those correlations that stand in certain relations to a third event deliberation in all relevant respects - i.e., those respects that make certain correlations exploitable and others not. But for this theory to be reductive, we should be able to capture the aspects of deliberation that make correlations exploitable or not in non-causal terms. So I need to show that the explanation of exploitable correlations proposed above doesn't rely essentially on the assumption that deliberation is a cause of action. That is, I need to show that the features of the relations between deliberation and options that explain why certain correlations are exploitable and others not can be characterized non causally.

And indeed, I think it can be shown that my explanation of exploitable correlations

doesn't rely essentially on the assumption that deliberation is a cause of action; instead, it hinges only on the following two non-causal conditions.

> **Correlation Condition**. For a to be a proper object of deliberation, the agent should regard her deliberation as correlated with the action. More specifically, the agent should believe that if she judges at the end of deliberation that a is preferable, she is more likely to do a than if she judges that $\sim a$ is preferable.
>
> **Temporal Condition**. For an action to be a proper object of deliberation, the action must lie in the *future* of deliberation.

For instance, in *Smoking*, smoking is a proper object of deliberation for Samantha because of the following two facts. First, her deliberation is correlated with whether she will smoke. In particular, if she judges at the end of her deliberation that smoking is preferable, she is more likely to smoke than if she doesn't. Second, her deliberation precedes whether or not she will smoke. I do not claim that these two conditions are sufficient for an action to be a proper object of deliberation. Rather, I claim that they are sufficient to explain why we can only exploit those correlations between act and outcome that correspond to the act causing the outcome: they can do all the job that the assumption that deliberation causes action played in the previous subsection.

Before seeing why let me make some brief comments on each condition. The correlation condition can be justified as follows. The goal of deliberating about what to do is in part to acquire information about what one will do. But for deliberation to fulfill this goal, the results of one's deliberation should be able to provide one with *information* regarding what one will do, which requires the results of one's deliberation to be correlated with what one will do. deliberating about a seems pointless. So there is a good justification for the requirement that an option for deliberation be correlated with the results of deliberation. The temporal condition encapsulates the fact that we can only deliberate about those actions that lie in the future of deliberation. Note that this temporal asymmetry of deliberation doesn't in itself prevent us from influencing the past. To see this consider Dummett's (1964) famous story of the dancing chief. Dummetts (1964) famous story of the dancing chief. Dummett imagines a tribe with the custom of sending its young people on lion hunts to prove their bravery. While they

are away, the chief performs dances intended to cause the hunters to act bravely. Surprisingly, the chief believes that dancing can have a retroactive influence on the hunters behavior, and thus continues to dance even after the hunt is over and the hunters are on their way home. (The chief justifies this policy by noting that on prior occasions where she danced after the hunt is over, it often turned out that the hunters acted bravely, while occasions where she stopped dancing after the end of the hunt often turned out to be accompanied by cowardice.) The time-asymmetry of deliberation doesn't prevent the chief's belief from being true. Even if the chief can only deliberate about whether he should dance in the future, this in itself doesn't prevent her contemplated actions to have effects rippling into the past. Indeed, on the view I proposed in the previous section, the dancing-bravery correlation will be exploitable as long as the chief (assumed to be a chance expert) regards her judgments about whether she should dance as correlated with the hunter's bravery, and this correlation doesn't disappear during deliberation. There is nothing logically impossible about this.

The temporal asymmetry of deliberation also plays a crucial role in Price's view of causation, although a different one than in my account. (I will compare the role that this asymmetry plays in Price's view and in mine in ch. 4, §4.4.) Like Price, I suspect that this asymmetry is a product of the fact that we are embedded in a world that displays certain pervasive physical asymmetries, in particular an asymmetry in the entropy gradient (see Price (2007, 278)[30]). But I do not have a more detailed explanation to offer. As we will see in the next chapter such an explanation isn't needed for my purposes, as the temporal condition doesn't figure in my metaphysical account of causal dependence. (Its only role is in explaining why we care about causal dependence in the context of practical reasoning.) The temporal condition isn't the only temporal asymmetry assumed in my explanation of exploitable correlations. Throughout this chapter, I assumed that deliberation itself is a process oriented in time, in the sense that the agent starts with credence and utility functions at a certain time to arrive at a judgment about what to do at a *later* time. This asymmetry can plausibly be traced back to the fact that deliberation is an *information-gathering process*, in which the

[30] Also Hartle (2005).

agent uses information about what she believes and desires at a certain time to arrive at a judgment about what she should do. So the temporal orientation of deliberation may be a consequence of a more general, *epistemic* time-asymmetry: the fact that we have a much stronger and secure epistemic access to the past than to the future. (In particular, we know much more about our beliefs and desires in the very recent past than about our future beliefs and desires.) This asymmetry is itself probably a product of the same facts about the initial conditions of the universe that explains the thermodynamic asymmetry. (See Albert (2000, ch. 6) for a detailed argument to this effect.) But here again, the exact origins of the temporal orientation of the deliberation process do not matter, as this asymmetry won't play a role in my metaphysics of causal dependence.

Let me now explain why I think the correlation and deliberation conditions are enough to explain why only those act-outcome correlations that correspond to the act causing the outcome can be exploited. Suppose that an event c causes another event e. Then in general, c will be uncorrelated with any event a that are temporally prior to e and correlated with it, unless there is a direct causal connection between c and a. For instance, consider smoking as a cause of lung cancer. One event that is prior to lung cancer and correlated with it is working in the asbestos factory. Here we won't find that smoking and working in an asbestos factory are correlated, unless there is a direct causal connection between them (e.g. working in an asbestos factory causes one to be a smoker).[31] Or consider driving drunk as a cause of road accident. One event correlated with road accidents and prior to it is snowfall. Here again, we won't find a correlation between driving drunk and snowfall, unless there is a direct causal connection between these two events (e.g. snowfall somehow causes people to drive drunk). This has the following consequence for us. Suppose that an agent is deliberating about an action a for the sake of an outcome o, where o is a cause of a. Since deliberation precedes a and is correlated with it, we should find that deliberation and o are uncorrelated, unless there is a direct causal connection between them. Suppose that deliberation

[31] Earlier I used this case as an example of the fact that causes of a joint effect are not correlated unless there is a direct causal connection between them. Here I am using it to describe a slightly different probabilistic pattern.

and o are uncorrelated, as in the original version of *Smoking*. Here we know that EDT won't recommend doing a for the sake of o. And if deliberation and o are directly causally connected (as in the variations on *Smoking* reviewed in §4.1), we know that the agent will during deliberation acquire evidence that screens off the correlation between deliberation and the outcome. So here again EDT won't recommend doing a for the sake of o.

Now consider the case where the o is an effect of the action a. In general, when c causes e, events prior to c and correlated with it are also correlated with e. Consider fever as a cause of shivering. Events prior to fever and correlated with it are correlated with shivering: if one gets the flu, or suffers from dehydration, one is more likely to shiver later. Or consider the switch being flipped as a cause of the light being on. Then events prior to the switch being flipped and correlated with it (Billy's desire to flip the switch, say) are also correlated with the light being on. From this and the two conditions I proposed, it follows when an act a causes an outcome o, we will find that deliberation about a is correlated with the occurrence of o. That is, if the agent judges that a is preferable, o is more likely to happen. But on the view I proposed in the earlier section this is exactly the sort of fact that makes a a good way to influence the occurrence of o.

Here is a worry for this argument. Suppose that an act is correlated with a desired outcome. And suppose moreover that the agent regards her deliberation as correlated with the act and the outcome. But suppose this is due not to the fact that deliberation causes the act, but to the existence of a common cause c of deliberation and the act. c causes the agent both to judge that a is preferable and (via a different route) to do a. The causal structure is represented in Figure 3.10. Here it is intuitively pointless for the agent to deliberate about the act, so that the act-outcome correlation cannot be exploited. But since now I am not relying anymore on the assumption that *bona fide* objects of deliberation must be caused by deliberation, my explanation entails that the act-outcome correlation is exploitable. The solution is to appeal one more time to the tickle defense, this time to show that during deliberation the agent will acquire evidence that screens off the correlation between her deliberation and her act. Suppose

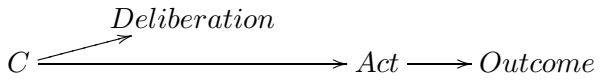

Figure 3.9

for simplicity that the common cause affects deliberation solely by influencing initial assessments of expected utility.[32] And suppose that a chance expert were to start deliberating about a and to discover that she initially prefers a. She would thereby evidence screening off the correlation between the remainder of her deliberation and the presence of the common cause. She would then know that if she were to deliberate further, her future judgments of utility would provide her with no information about what she will eventually do. Moreover, even before starting to deliberate the agent can predict all of this: she knows that she is guaranteed to acquire evidence given which the correlation between deliberation and a disappears. So we have a simple explanation for why the agent shouldn't deliberate about a: she can recognize at the outset that she will come to regard deliberation as pointless. So we can escape the problem by strengthening a little bit the correlation requirement. In addition to requiring deliberation to be correlated with action, we should require the correlation to be robust under the evidence acquired during deliberation itself. At no point during deliberation should the agent acquire evidence given which the remainder of her deliberation provides her with no information about what she will do.

I conclude that properly interpreted the temporal and correlation conditions are enough to explain why only those act-outcome correlations that correspond to the act causing the outcome can be exploited. In particular, given these two conditions we can explain the asymmetry in exploitability between causes and effects as the product of the following asymmetry: when c causes e, then e will be correlated with events preceding c and correlated with c. But c will not be correlated with events prior to e and correlated with it, unless there is a direct causal connection between them. The asymmetry in exploitability between causes and effects can then be recast as a consequence of the fact that deliberation must be correlated with action and precede it; and thus that it will

[32] As will hopefully be obvious by now, the reasoning generalizes to more complex cases in which the common cause also affects further judgments.

be correlated with the action's effects, but not its causes.[33] This result will pave the way for the account of causal dependence I will offer in the next chapter.

[33]Unless there is a direct causal connection between these causes and the deliberation, but any such direct connection will be screened off during deliberation itself.

Chapter 4

Probabilistic Interventionism

In the previous chapter, I argued that EDT provides a good account of what makes a correlation exploitable or not. In particular, EDT entails that an act is a good way to usefully influence a desired outcome only when the outcome is intuitively causally dependent on the act. As this account relies on the evidential principle of rational choice, which is in itself very plausible, it thus constitutes a good *explanation* of the practical relevance of causation. So a theory of causal dependence that bears a close connection to this explanation would be guaranteed not to make the practical relevance of causation a mystery, by contrast to the theories I reviewed in chapter 2. The simpler way to develop such a theory would be to identify causal dependence with those relations that are exploitable according to EDT. But as I already pointed out, this is too quick. EDT is solely concerned with relations between *human actions* and desired outcomes. But causes need not be human actions. Indeed, some causes (such as the position of the moon) are simply not the sort of thing that we can manipulate. My suggestion was to adopt the interventionist strategy and characterize e's causal dependence on c in terms of the relations between c, e and a third event. This third event should be like deliberation in all relevant respects - that is, it should have the properties that play a crucial role in explaining what makes act-outcome correlations that we regard as causal exploitable and other correlations unexploitable. The intent is that this third event need not be a human deliberation or anything agentive, so that c need not be a human action to be a cause. I will call this third event a *probabilistic intervention*

or *p-intervention* for short. The goal of this chapter is to articulate the notion of a p-intervention (§1) and to extract an account of causal dependence on its basis (§2). I will call this account *probabilistic interventionism*. In §3, I will argue that probabilistic interventionism provides a good solution to Russell's problem, and in §4 I will briefly compare this account to other solutions to Russell's challenge.

1 P-Interventions

The intent behind the notion of a p-intervention is that a p-intervention on c should have all the relevant properties of deliberating about an action a - i.e., the properties that play a crucial role in explaining why an action a can be used to influence those events we intuitively regard as its effects but not those we regard as its causes. As we might also put it, the idea is to define the notion of a p-intervention in the image of deliberation.

Now, according to the explanation I gave in chapter 3, §5, there are two properties of deliberation which play a particularly important role in explaining why we can influence effects but not causes of our actions. These are the correlation and temporal conditions respectively. Deliberation should be correlated with action, and it should precede action. The first step, then, is to require a p-intervention on c to have the same two properties with respect to c. Thus, a p-intervention on c (denoted i_c) should be an event temporally prior to c, and correlated with c. It will be useful to get clear on what the latter clause means exactly. In the case of deliberation, what I meant when I say that deliberation is correlated with an action a is that the agent is more likely to do a if she judges that a is preferable than if she judges otherwise. Likewise, for c to be correlated with i_c is for c to be more likely to happen given i_c than given some alternative, incompatible event $i_{\sim c}$. We might call $i_{\sim c}$ the *intervention contrast*. In turn, this correlation claim should be read as follows: there are reference classes for c, i_c and $i_{\sim c}$, C, I_C and $I_{\sim C}$ respectively, such that the general probability of C is higher given I_C than given $I_{\sim C}$.

As we also saw in ch. 3, there is another feature of the deliberation process that is

absolutely crucial to explaining why an act can be used to influence its effects but not its causes. This is the fact that while deliberating about an action, the agent regards the outcome of her deliberation process as evidentially independent of those events that are intuitively *causes* of her action. (Or at least, she will come to regard them as evidentially independent conditional on the evidence she acquires during deliberation.) For instance, in *Smoking*, Samantha regards her possible judgments regarding the preferability of smoking as evidentially irrelevant to the presence of the chemical, as the two are uncorrelated. And in variations on *Smoking* in which deliberation and the chemical are correlated, Samantha will during deliberation acquire information screening off the possible results of her deliberation process from the presence of the chemical. This feature of Samantha's deliberation process was absolutely crucial, since it allowed us to give an evidentialist explanation for why the smoking-chemical correlation is not exploitable. A p-intervention should mimic this feature of the deliberation process. That is, a p-intervention on c should be the outcome of a process whose possible outcomes are uncorrelated with those events that we intuitively regard as causes of c. But of course we cannot simply *stipulate* that the p-intervention be independent of the causes of c. Otherwise the notion of a p-intervention wouldn't be reductive.

My proposal is to require a p-intervention and its contrast to be alternative possible outcomes of a *random process*. More precisely, the reference classes I_C and $I_{\sim C}$ should be possible alternative kinds of outcomes of a random process. In the first instance, by 'random process' I mean processes like the toss of a fair coin, the spinning of a roulette wheel, the operation of a random number generator, and so on. On this proposal, a p-intervention on (e.g.) Samantha smoking might be the following event. Suppose that a fair coin is tossed, and that Samantha is more likely to *try* to smoke if the coin lands heads than if it lands tails, so that in turn if the coin lands heads Samantha is more likely to smoke than if it lands tails. We assume that the outcome of the coin toss takes place before Samantha decides to smoke. The coin landing heads counts as a p-intervention on Samantha smoking, where the intervention contrast is the coin landing tails. (Here the relevant reference classes are FAIR COIN LANDING HEADS, FAIR COIN LANDING TAILS and SAMANTHA SMOKING.) The point of requiring

a p-intervention on c and the intervention contrast to be outcomes of a random process is that, plausibly, they will be independent of those events that we intuitively regard as causes of c. In the example just described, if the coin toss process is random we should observe no correlation between heads (contrasted with tails) and the presence of the chemical. That is, those times on which the coin lands heads should not be any more associated with the presence of the chemical than those times on which the coin lands tails. As we might also put it: when the chemical is present, if the coin toss process really is random the coin is as likely to land heads as it is to land tails; *and likewise when the chemical is absent.* If the coin toss really is random the presence or absence of the chemical should not be correlated with its outcome.[1] By contrast the outcome will be correlated with those events that we intuitively regard as effects of Samantha smoking. For instance, those times on which the coin lands heads should be followed more often by Samantha smoking than those times on which the coin lands tails, and hence they will also be more often followed by the presence of nicotine in Samantha's blood.

At this point obviously I need to say more about the notion of a random process. I think the notion can be cashed out via the work done by Strevens (2011)[2] on how non-trivial probabilities can emerge from a deterministic dynamics. Consider the toss of a fair coin again, which is a process that we regard as paradigmatically random. Now take all the possible microstates that intantiate a toss of a fair coin. This is the set of possible initial microconditions of the process. All of these microconditions correspond to a toss with a certain initial upward velocity v of the coin and a certain angular velocity w of its spinning. Now consider the physical dynamics of the coin toss, assuming that the coin begins with the heads side up. The dynamics can be represented in the following figure. (see Keller (1986)). Let the x axis represents the initial upward

[1] As we will see in the next section, this isn't entirely true. Intuitively the correlation between the outcome of the coin toss and the presence of the chemical cannot arise because the latter causes the coin to land heads rather than tails. If the process really is random the chemical cannot have such an influence on its outcome. But a correlation might arise because the outcome of the coin toss somehow causes the presence of the chemical. So the appeal to the notion of random process isn't enough to ensure that a p-intervention on c will not be 'directly causally connected' with the causes of c, as it should be if it is to mirror deliberation. I will deal with this problem in the next section.

[2] Based in part on the work of Keller and Diaconis. See (Keller, 1986) and (Diaconis, 1998).

velocity of the coin and the y axis its angular spinning. Let black points represent initial microconditions that evolve into a heads outcome while white points represent initial microconditions that evolve into a tails outcome. As Strevens points out, the ratio of the black sections to the white sections is the same in every small neighborhood of the space of possible initial micro-conditions, namely 1 to 1. In other words, for any small volume of possible initial upward velocities and angular spinnings, the proportion of conditions that lead to the coin landing heads is $\frac{1}{2}$ or very close to $\frac{1}{2}$. Strevens calls this property of the dynamics *microconstancy*. That the dynamics is microconstant means in particular that for any exact initial upward velocity (or angular spinning) that leads to the coin landing heads, a very small change in its value would lead to the coin landing tails instead. In that respect, the outcome of the toss of a fair coin depends *extremely finely* on its exact initial microcrondition. Strevens goes on to show that in virtue of the microconstancy of the dynamics, every reasonably smooth probability distribution over the initial conditions of the coin toss will nomologically evolve into a probability $\frac{1}{2}$ of the coin landing heads. On Strevens's view this is what grounds the fact that the coin has a $\frac{1}{2}$ chance of landing heads.

My proposal is to identify random processes through the property of microconstancy, as follows. Let's say that a random process is the trajectory in spacetime of a certain object (e.g. a coin) with the following characteristics. The direction of the trajectory is stipulated to be from past-to-future. (Remember that we are using the notion of a random process with the intent of mimicking certain properties of deliberation; here you might think of this temporal condition as mimicking the fact that deliberation is a process oriented from past to future.) At the earlier endpoint of the trajectory, a certain macroscopic event happens to the object (e.g. the coin is tossed). At the later endpoint of the trajectory, the object is in two alternative possible states (e.g. being heads up or tails up). These are the *outcomes* of the process. The last and crucial requirement is that the physical dynamics leading from the earlier endpoint to the later one be microconstant. That is, in any small region of the space of possible initial microconditions that realize the microscopic event, the ratio of conditions that produce one outcome to conditions that produce the other outcome should be 1 to 1. This characterization of

random processes makes no explicit reference to causal facts; it appeals only to facts about the microphysical dynamics of the process (microconstancy) and a stipulation about the direction of time in which the dynamics should be microconstant. The idea is that a p-intervention and its contrast should be alternative outcomes of a random process in this way of cashing out the notion of a random process.

To summarize, we can define the notion of a p-intervention as follows:

> **P-Intervention.** Take an event c. Then an event i_c is a p-intervention on c and $i_{\sim c}$ its intervention contrast just in case
>
> 1. i_c and $i_{\sim c}$ are temporally prior to c;
> 2. i_c and $i_{\sim c}$ are alternative outcomes of a random process;
> 3. c is more likely to happen given i_c than given $i_{\sim c}$. That is, $P(c/i_c) > P(c/i_{\sim c})$.

Although the notion of a p-intervention with the intent of mirroring rational deliberation, it makes no reference to human agency or intention. Consequently, a p-intervention need not involve any human action or intention at any point. For instance, consider the event of the barometer having a high reading (rather than a low one). Consider for instance a mechanical device that first randomly outputs a '1' or a '0'. If the machine outputs 1, it sets the barometer at a high reading; if it outputs 0, it sets the barometer at a low reading. We allow for circumstances in which the machine fails, so that a 1 isn't necessarily followed by a high reading on the barometer. All that we require is that the barometer be more likely to display a high reading on those times where the machine outputs 1 than on those times at which the machine outputs 0. Then the output 1 constitutes a p-intervention on the barometer having a high reading, and the output 0 constitutes its intervention-contrast. This shows that a purely natural process can constitute a p-intervention on c, as long as it stands in the right probabilistic and temporal relations to c. This will be important for the next section, as it will allow us to formulate a theory of causal dependence that both preserves a tight connection with the theory of exploitable correlations offered in ch. 3 and correctly handles cases involving causes that are not human actions.

2 Probabilistic Interventionism

Using the concept of a p-intervention, we are now in a position to formulate a new account of what it is for an event e to causally depend on c. The guiding idea is that for e to causally depend on c, e should stand in the same relations to a p-intervention on c as an outcome o stands to a deliberation about an act a when the a-o correlation is exploitable.

Here is a natural preliminary suggestion motivated by the analysis of exploitable correlations offered in ch. 3. There we saw that the crucial feature that makes (e.g.) the smoking-chemical correlation *un*exploitable is that the results of Samantha's deliberation about whether to smoke and the presence of the chemical are uncorrelated with each other, despite the fact that both are correlated with smoking. Thus the suggestion is to require, similarly, that a p-intervention on c be correlated with the occurrence of e for e to causally depend on c. In other words, the idea is that e causally depends on c only if $P(e/i_c) > P(e/i_{\sim c})$. According to this proposal, what makes it so that atmospheric pressure isn't causally dependent on the reading of the barometer is this. Take a p-intervention on the barometer having a high reading. For instance, this might be the output 1 of the random device described at the end of previous section. Then the output will not be correlated with the atmospheric pressure being high (rather than low). That is, those times on which the random device outputs 1 won't be associated with high atmospheric pressure than those times on which the random device outputs 0. By hypothesis, the outcome of the random device depends extremely finely on its initial microcondition. For there to be a correlation between its output and atmospheric pressure, there would have to be a strange correlation between the exact initial microconditions of the machine and the presence of high atmospheric pressure. By contrast, the condition I just proposed correctly allows the throwing of a rock to be a difference-maker for the window breaking. To see this, suppose that Suzy tosses a fair coin. If the coin lands heads, she will try to throw the rock, whereas if it lands tails, she will try not to throw the rock. The coin landing heads constitutes a p-intervention on the rock being thrown. And here we will find a correlation between the coin landing heads

and the window breaking. Times on which the coin lands heads are more likely to be followed by Suzy throwing the rock, and thus more likely to be followed by the window breaking shortly afterward.

So this proposal seems to be on the right track. However, it is not entirely right. Reconsider the barometer example I just gave. Intuitively, for there to be a correlation between the p-intervention on the barometer and atmospheric pressure, there would have to be a direct causal connection between the two. One would have to cause the other, or there should be a common cause. And the idea was that such a correlation could not be due to the atmospheric pressure causing the random device to issue output 1 rather than 0 (or to a common cause of pressure and of the outcome). For this to be the case, there would have to be a way for atmospheric pressure or the common cause to 'select' an initial micro-condition evolving into output 1. This would require incredibly fine selection powers on the part of pressure or the common cause, as such initial micro-conditions are scattered throughout the space of possible initial conditions for the random process, and each one is neighbored by initial micro-conditions evolving into output 0. So no correlation between the p-intervention and a cause of the barometer reading can arise from a direct causal influence of the barometer's cause on the p-intervention. But there might still be a correlation due to causal influence going the other way around, from the p-intervention to the cause. More generally, a p-intervention on c might well be correlated with an event that we intuitively regard as a cause of c if the p-intervention has a direct causal influence on this event. Here is a case to illustrate. consider again the random device whose outputs are correlated with the barometer having a high or low reading. And suppose that if Billy sees the machine outputting 1, he will find a nearby window and throw a rock at it. (If the machine outputs 0, he will do nothing.) The causal structure of the case is represented in Figure 4.2. There we will find that on those times when the machine outputs 1 are correlated with a high reading on the barometer with a nearby window shattering, but of course the high reading on the barometer doesn't cause windows to break. To put it intuitively, we need to find a way to require a p-intervention on c to be correlated with a putative effect e only (if at all) through its correlation with c.

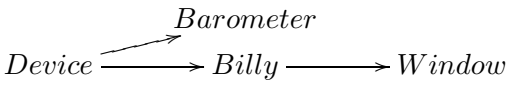

Figure 4.2

Since p-interventions are supposed to mirror deliberations, to solve the problem it is a good idea to look at what happens when deliberation about a has a direct influence on one of a's causes. We already encountered such a case in ch. 3, §4.1. There we considered a variation on *Smoking* in which Samantha's initial utility judgment somehow influences the presence of the chemical. There we saw that by updating on this judgment she acquires evidence that screens off the correlation due to this causal influence. In other words, the correlation disappears during stages of her deliberation that are temporally intermediary between her initial utility judgment and her action. Similarly, then, we should allow e to causally depend on c only if stages temporally intermediary between i_c and c are also correlated with e. This can be made more precise as follows. Let's say that an event mediates the correlation between i_c and c when it screens off their correlation and is temporally intermediary between them. Suppose for instance that after the device has outputted '1' or '0', it manipulates the barometer dial through a mechanical arm. Then the position of the mechanical arm mediates the correlation between the random outputs of the device and the barometer's reading. Hold the position of the arm, and the output of the device is independent of the barometer reading. We can then require effects of the barometer's reading to be correlated with the position of the arm, holding fixed the output of the device. Since holding fixed the output of the device, the position of the arm is uncorrelated with the window breaking, we get the result that the window breaking isn't an effect of the barometer reading.[3] More generally, the correlation between e and i_c should be robust in the sense that for any mediating event m, holding fixed i_c and other mediating events earlier than c, m is still correlated with e. This reflects the fact that at any stage of her

[3]The position of the arm might itself be a direct cause of the window breaking, in which case it will still be correlated with it. There the idea is that one will still be able to find a temporally later mediating event that is uncorrelated with the window breaking given previous mediating events. But what if every event mediating the output of the device and the barometer reading is causally relevant to the window breaking? Here my answer is that such cases are sufficiently fanciful and unrealistic to be properly ignored in a theory concerned with explaining causation in the actual world.

deliberation, an agent should regard her deliberation about a as correlated with o for a to be a good way to influence o. Specifically, there should still be a correlation between any stage of her deliberation and the occurrence of o, holding fixed the previous stages of her deliberation.

We have one last refinement to make. Consider again the case where Suzy throws a rock at a window. Independently, Billy also throws a rock at the window; he is an excellent thrower, so that his throw is guaranteed to break the window. In those circumstances Suzy's throwing a rock is not a difference-maker for the window breaking. But a p-intervention on Suzy throwing the rock will nonetheless correlates with the window breaking. Consider for instance Suzy deciding to throw the rock (rather than deciding not to), which (let's assume) qualifies as a p-intervention on her throwing the rock. Then the probability of the window breaking is higher given Suzy's deciding to throw the rock than given Suzy's deciding not to. Let 'window' denote the window breaking, 'decision' denote Suzy deciding to throw the rock, and 'billy' denote the presence of Billy. If it is the outcome of a random process Suzy's decision will be uncorrelated with whether Billy throws his rock. So we have $P(window/decision) = P(window/decision.billy)P(billy) + P(window/decision.{\sim}billy)P({\sim}billy)$. Now given the presence of Billy Suzy's decision doesn't correlate with the window breaking; but given Billy's absence it does. So on the whole Suzy's decision will correlate with the window breaking. What we need is to find a way to hold the presence of Billy fixed when assessing whether Suzy's throw makes a difference to the window breaking. Here is a natural way to do so. When assessing whether a p-intervention on Suzy throwing the rock is correlated with Billy throwing, we will find no such correlation (or no correlation that is screened-off when the correlation between the p-intervention and Suzy throwing the rock is screened-off. This allows us to determine that the presence or absence of Billy isn't the sort of thing that Suzy throwing the rock can influence. More generally, testing for correlations between p-interventions on c and another event d allows us to determine whether d is the sort of event that c could be used to influence. That is, it allows us to determine which other events a certain event can *in principle* influence. We can then require an event e to be actually causally dependent on c just in case, holding

fixed all those actual events that c cannot influence, e is correlated with c. This means that in the case at hand, when checking whether the window breaking causally depends on Suzy's throw, we should hold fixed whether or not Billy independently throws a rock. In circumstances where he does, we won't find a correlation between Suzy's throw and the window breaking.

We thus arrive at the following account of causal dependence. I will call it *probabilistic interventionism* (PI), since its main tool is the notion of a p-intervention. It starts with a definition of an event being able to influence another, and then uses this notion to define causal dependence.

> **PI.** Take an event c, and let i_c be a p-intervention on c, with an intervention contrast $i_{\sim c}$. Then c can influence an event d just in case
>
> 1. i_c is correlated with d. That is, $P(d/i_c) > P(d/i_{\sim c})$.
> 2. The correlation is robust: any mediating event between i_c and c is also correlated with d, even holding fixed temporally earlier mediating events.
>
> Now take two events c and e. e causally depends on c just in case $P(e/c.\mathbf{k}) > P(e/{\sim}.\mathbf{k})$, where $\mathbf{k}$ is the set of all events that c cannot influence.

3 Probabilistic Interventionism as a Solution to Russell's Problem

My goal in this dissertation was to provide a new solution to Russell's problem. In this section, I will show how probabilistic interventionism answers Russell's challenge of finding a place for causation in the physical world, using as benchmark the criterion of evaluation I proposed in chapter 1. In the next section, I will offer some points of comparison between my solution and the other accounts of causal dependence surveyed in previous chapters.

A proper solution to Russell's problem should show how causal facts emerge from the fundamental physical structure of the world, which is not itself causal. On the view I proposed, all the materials that come into play in explaining facts about causal dependence are themselves straightforwardly grounded in fundamental physics. **PI**

appeals to the notion of a p-intervention, and requires cause and effect to stand in certain probabilistic relations to a p-intervention. The relevant probabilities here are statistical-mechanical objective probabilities interpreted in a Humean fashion, along the lines of ch. 2, §2.2. As we saw there, the existence of these probabilities is clearly compatible with fundamental physics, and can be explained using only material one finds in fundamental physics. More precisely, they are generated by taking the standard Lebesgue measure over all possible microstates of the universe consistent with the PH, and evolving the measure forward or backward in time through the physical laws. These probabilities earn their name of 'chance' by figuring in the best systematization of the pattern of actual physical events.[4] The notion of a p-intervention itself appeals to certain facts about those probabilities, as well as to the notion of a random process. In §1 of the present chapter, I argued that the notion of a random process can be cashed out in terms of the microconstancy of the physical dynamics (together with a stipulation that the dynamics must be microconstant in the forward direction of time). Here again, no material over and above the one contained in fundamental physics is needed. Thus **PI** deserves to be called a *physicalist* theory of causal dependence.

Moreover, **PI** doesn't lead to eliminativism, by contrast to the view endorsed by Russell on which causation is localized, asymmetric physical determination. Indeed, **PI** explains how fundamental physics leaves space for localized, asymmetric causation. Consider first the problem of the compatibility of localized causation with the globality of physical laws. The solution that **PI** offers is the same as the one proposed by all probabilistic accounts of causal dependence.[5] On **PI**, for c to cause e is fundamentally for c to be correlated with a certain event that is correlated with e. Insofar as correlation doesn't require physical determination, **PI** thus allows localized events to count as causes. A p-intervention on throwing a rock can be correlated in the right way with a window breaking so that throwing a rock counts as a cause of the window breaking on **PI**, even if the rock being thrown doesn't physically determine the window to break.

[4] Moreover, these probabilities are compatible both with deterministic and indeterministic fundamental physical theories. More generally, my account doesn't rely on the details of the fundamental dynamics of our world, but on very general features of it. In that respect, it satisfies the virtue of generality discussed in ch. 1, §4.2.2.

[5] See chapter 2, §3.

Consider next the problem of the compatibility between directed, time-asymmetric causation and the symmetry of fundamental physical laws. Here **PI** ticks the right boxes: it entails both that causation must in our world be predominantly past-to-future, while leaving a loophole for backward causation in exceptional circumstances (such as those involving time-travel).

To see this, note first that in itself **PI** doesn't make backward causation impossible. Consider again Dummett's (1964) case of the dancing chief introduced in the previous chapter, where the bravery of the hunters at a certain time depends on the chief dancing in the *future*. **PI** allows this to be the case, along as the following probabilistic facts hold. Suppose that the chief decides whether to dance by throwing a fair coin. If the coin lands heads, the chief decides to dance; if it lands tails she decides not to. Then the coin landing heads constitutes a p-intervention on the chief dancing. **PI** will entail that the dance causes the bravery if the coin landing heads is correlated with the hunters' being brave.[6] In itself this is perfectly possible. There is no contradiction in postulating a correlation between the outcome of the coin toss and the hunters' past behavior. But this case also shows why on **PI**, backward causation would constitute a truly exceptional circumstance. Although perfectly logically possible, Dummett's case instantiates a probabilistic pattern that we never observe in our world. In Dummett's case, there is a correlation between the hunter's behavior and the microcondition in which the coin toss starts: bravery is correlated with the coin starting in a heads-conducive micro-condition. And intuitively, this correlation cannot be due to the influence of one on the other. There is no physical mechanism by which the initial microcondition of the coin could 'send a signal' telling the hunters to be brave. Likewise, there is no physical mechanism by which the hunter's bravery can 'select' an initial head-conducive micro-condition. (Such a selection process would have to be incredibly sensitive, as the heads-conducive initial microconditions are scattered throughout the space of initial conditions, and always neighbored by tails-conducive initial microconditions.) So the correlation can only be traced back to the fact that the coin landing heads and the hunter's bravery are both correlated with an event in their common *future* - the chief's

[6] And some other screening-off conditions are satisfied, but for simplicity we can leave those aside here.

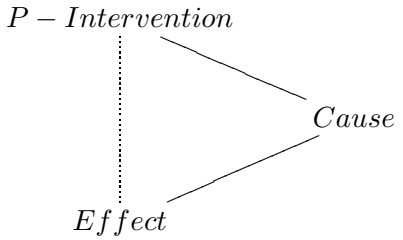

Figure 4.3

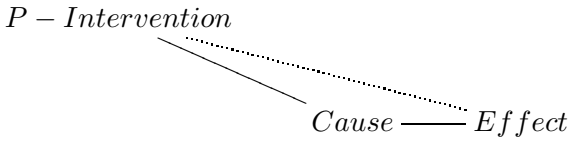

Figure 4.4

dance. As we might also put it, the dance must induce a correlation between two earlier events. More generally, a case of backward causation must on **PI** involve the following probabilistic pattern: two events (the p-intervention and the effect) being correlated with each other, in a way such that the correlation can only be traced back to their joint associations with an event in their common future (the cause). This probabilistic structure is represented in Figure 4.3, where the dotted line indicates that the correlation between the p-intervention is due to their joint associations with the cause. But as we saw when discussing the fork asymmetry in ch. 2, §3, this probabilistic pattern (corresponding to a fork open to the past) would be truly exceptional. It never occurs in our world, and as I noted in ch. 2 this is probably because of those features of the SM-probability distribution that also explain the thermodynamic asymmetry. Thus, on **PI**, the exceptional nature of backward causation follows from the fact that a case of backward causation would have to constitute a probabilistic pattern that we never observe in our world.

Conversely, on **PI**, forward causation is an entirely routine occurrence in our world because cases of forward causation must instantiate a probabilistic pattern that is itself very common in our world. A case of forward causation must involve a correlation between a random process and an event (the effect e) in its future, such that the correlation goes by way of a temporally intermediary event (the cause c), as in Figure 4.4. Such probabilistic patterns are entirely ordinary occurrences, and there is nothing surprising to them. For instance, suppose that Billy throws a rock at a window if a tossed coin lands heads, and does nothing otherwise. Here we will observe a correlation

between the coin landing heads and the window breaking later generated by the joint associations of these events with the temporally intermediary event of Billy throwing the rock.

Note that the contrast between the rarity of probabilistic structures like in 4.3 and the commonality of structures like in 4.4 constitutes a temporal asymmetry, one slightly different from the fork asymmetry. The asymmetry is as follows. Suppose that a p-intervention is correlated with an event c in its future. Then the p-intervention is often correlated with other events e that are themselves in the *future* of c (4.6). But it is never correlated with other events that are themselves in the *past* of c. This probabilistic asymmetry bears a strong resemblance to the fork asymmetry. Both asymmetries are constituted in part by the rarity of past-directed open forks. But while the fork asymmetry contrasts this with the commonality of future-directed open forks, the asymmetry under consideration contrasts it with the commonality of a different kind of probabilistic pattern, in which a correlation between two events is induced by an event temporally intermediary between them. On **PI**, the time-asymmetry of causation depends on this pervasive statistical asymmetry, not on any asymmetry in the fundamental physical laws themselves. **PI** thus solves Russell's challenge of finding a physical basis for the direction and time-asymmetry of causation.

Most importantly, **PI** provides a good explanation of effective strategies. In particular, **PI** together with the view proposed in chapter 3 give us a good explanation of why we should care about causal dependence in the context of decision-making. On **PI**, e's causal dependence on c is a matter of their relations to a p-intervention on c, which has all the relevant properties of a deliberation. We are interested in those relations because they are exactly the sort of relations that should hold between act, outcome and deliberation when the act is a good way to influence the outcome according to EDT. Insofar as EDT is a plausible theory of rational choice, we thereby have a good explanation for why we should care about causation in the context of rational decision-making.

4 Comparison with Other Accounts

By way of conclusion, I will briefly compare **PI** with other accounts of causal dependence discussed earlier, with an emphasis on showing how **PI** solves or escapes the problems into which other accounts run.[7] One respect in which **PI** works better than other accounts I discussed should by now be clear. In chapter 2 and elsewhere, I argued that leading solutions to Russell's problem all make it in some way or other mysterious why causal dependence should be the relation that matters for rational decision-making. By contrast, I have argued, there is a good evidentialist explanation for why the relation encoded in **PI** is the one we should be concerned about in the context of assessing what actions it would be in our best interest to do.

4.1 The Temporal Theory of the Causal Direction

Let's start with the temporal theory of the causal direction, discussed in ch. 1, §4.2.3. One similarity between the temporal theory of the causal direction and **PI** is that both appeal to the past-future direction to ground facts about causal dependence. The temporal theory does so by requiring causes to come earlier than their effects, while mine does so by requiring p-interventions to come earlier than events p-intervened upon. (This mirrors the fact that we can only deliberate about actions in our future, not actions in our past, and in that respect the appeal to the temporal direction incorporated in **PI** isn't arbitrary.) My theory, however, doesn't rule out the possibility of backward causation in exceptional circumstances, as we saw in the preceding section. Another advantage of **PI** over the temporal theory is that it is more unificatory. There are good reasons to think that the statistical asymmetry which on **PI** underlies the time-asymmetry of causation has its roots in the same fact as the thermodynamic asymmetry. In that respect, **PI** has a better chance of leading to a unified account of physical asymmetries than the temporal theory.

[7] I leave aside process theories.

4.2 Lewis's and AKL's Accounts of Causal Dependence

One respect in which my theory is very similar to Lewis's and AKL's accounts of causal dependence is as follows. On all of those accounts, to evaluate e's causal dependence on c, we should compare situations in which c happens and in which it doesn't that are otherwise very similar to actuality. According to **PI**, when evaluating causal dependence we should consider situations in which all those actual events that c cannot influence happen. **PI** also borrows a major element from AKL's account, namely its understanding of objective probabilities as statistical-mechanical Humean chances. But **PI** escapes the problems that these two accounts face. One such problem had to do with Lewis's and AKL's explanations of the time-asymmetry of causation. Our world doesn't display Lewis's asymmetry of miracles; and the statistical asymmetry of records on which AKL rely may be too weak to make the causal asymmetry as strict as we think it is. By contrast the statistical asymmetry which underlies the time-asymmetry of causation on **PI** exists, and is much stronger than the asymmetry of records. Whereas it is relatively easy to imagine ordinary situations in which there are no or few records of a past event, circumstances in which an event is correlated with the outcome of a random process in virtue of their joint associations with a third event in their common future would be truly exceptional. Another advantage of **PI** over AKL's account is that it doesn't make simultaneous causation impossible *a priori*, and it doesn't rely on a libertarian conception of control. Insofar as my account embodies a theory of control, it is the idea that our deliberations are (at least sometimes) correlated with our actions. This is perfectly compatible with the physical dynamics of our world, whether or not this dynamics is deterministic. Finally, my account gives the right results in Frisch's case of the piano piece, which is a counterexample to AKL's theory. **AKL**, remember, seems to entail that my decision to play the second ending is a way to causally influence my playing an earlier part of the piece, whereas the causal relation goes the other way around. Now consider a p-intervention on my decision to play the second ending. Suppose for instance that if a coin lands heads I decide to play it, whereas if the coin lands tails I decide not to. There we will find that despite being both correlated with my playing the second ending, the outcome of the coin toss and my playing the earlier

part will be uncorrelated with each other. It would be extraordinary if the coin had more chance of landing heads than tails on those times where I have played the earlier part already. Consequently **PI** rightly entails that my playing the second part isn't a way to influence what I played earlier. This reflects the fact that in the context of deliberating about whether to play the second ending, a chance expert would come to regard her judgments about what to do as evidentially irrelevant to whether she played the earlier part.

4.3 Reichenbach's and Papineau's Accounts of Causal Dependence

PI also bears strong similarities to Reichenbach's and Papineau's probabilistic account of causal dependence. In particular, like them it relies on the idea that to distinguish between cause and effect, we should look at the probabilistic relations between cause, effect and other events - i.e., at the places of the cause and the effect within a probabilistic network. Here the relevant event is a p-intervention: by taking into account c's and e's places in a probabilistic network including a p-intervention on c, we can determine whether c is the sort of event that can be used to influence e. Moreover, **PI** relies on many of the principles at the center of the probabilistic approach, in particular the causal Markov condition (in the form of the assumption that whenever two events are correlated, one causes the other or they have a common cause). Finally, the account of the time-asymmetry of causation I offered is in certain respects similar to the one offered by Reichenbach and Papineau. In particular, on my account the exceptional character of backward causation is due to the exceptional character of forks open to the past whose prongs involve the outcome of a random event. On the view I proposed, however, the routine character of forward causation isn't explained by the abundance of forks open to the future. Rather, it is tied to the fact that it is very common for the outcome of a random process that is correlated with an event c in its future to also be correlated with events in c's future itself.

PI escapes the problems raised earlier for Reichenbach and Papineau. By contrast to Reichenbach's account, **PI** doesn't make unconditional correlation necessary for causal dependence. Rather, effect and cause should be correlated holding fixed all those actual

events that the cause cannot influence. In the case proposed by Cartwright involving Simpson's paradox (see Fig. 2.3 in ch. 2, §3.1), this will include the presence or absence of the gene, conditional on which heart attacks and smoking are positively correlated. Turning to Papineau's account, **PI** also escapes the problems involving the faithfulness condition. Consider again Hitchcock's example of an unfaithful causal structure, in which smoking and lung cancer are uncorrelated because a common cause (leaving in the countryside) masks their correlation. (See Figure 2.13 in ch. 2, §3.2.1.) On **PI**, whether (e.g.) John lives in the countryside will be one of the factors to be held fixed when assessing whether John's cancer is dependent on him smoking. Consequently, **PI** will rightly entail that it does. Another problem with Papineau's account had to do with his assumption that for any causally ambiguous statistical structure, taking into account more variables will always allow us to determine which causal hypothesis is correct. As we saw, Papineau does little to motivate this assumption. My account relies on a similar assumption but is more motivated, as I have shown how to construct a variable (the p-intervention) whose probabilistic relations with c, e and other events allows us to determine whether e causally depends on c.

4.4 Price's Account

Finally, let me compare **PI** with Price's account of causal dependence. We saw that both accounts rely on a similar guiding idea, namely that what makes a correlation exploitable or not is to be explained in evidential decision-theoretic terms. I argued that Price's way of doing so runs into problems, however, due to the fact that the tickle defense relies on the assumption of control. I proposed a different evidential explanation of exploitable correlations that doesn't rely on this assumption. Another similarity between the two accounts is that both rely on the fact that we can only deliberate about actions in our future. This time-asymmetry of deliberation plays different roles in Price's account and mine, however. On my view, the time-asymmetry of deliberation plays no role in the metaphysics of causation itself. It only comes into play in explaining why we care about causal dependence in the context of rational decision-making. Specifically, it explains why we care about the relation between a

cause, its effect and a certain event temporally prior to the cause (the p-intervention). We care about them in part because our own deliberations are temporally prior to the actions they are about. On Price's view, the time-asymmetry of deliberation comes into play in explaining why causes precede their effects and why we cannot influence the past (Price and Weslake, 2009). However, by itself the time-asymmetry of deliberation isn't sufficient to explain why we cannot influence the past. To see this, consider Dummett's story of the dancing chief again. There the chief can influence the hunters' past behavior by dancing, even though the object of her deliberation (dancing) lies in the future of her deliberation. This is enough to show that by itself the time-asymmetry of deliberation doesn't prevent us from influencing the past. Indeed Price himself concedes the point when he and Weslake point out that the time-asymmetry of deliberation 'leaves 'loopholes for exceptional cases' (435) in which one can influence the past. (By their own lights, Dummett's example s such an exceptional case, because the correlation between dancing and the hunter's past behavior doesn't disappear when the chief deliberates about dancing.) The upshot is that Price's view leaves us with a mystery. If the time-asymmetry of deliberation by itself doesn't prevent us from influencing the past, why are such cases at best exceptional in our world? And why, by contrast, are cases in which we can influence the future so common? A natural thought to have here is that this must be in some way the result of an objective, statistical asymmetry. But Price does not tell us what this asymmetry is. Indeed, he suggests that no statistical asymmetry can come into play in explaining why we cannot influence the past.[8] The view I proposed solves this mystery: the time-asymmetry of causation and influence is a product of the statistical asymmetry described in §3 of this chapter.

There is another difference between Price's account and the one I proposed here. In his (2007), Price argues that the only solution to Russell's challenge is a form of *perspectivalism* or *projectivism* about causation, on which causation is a projection of our perspective as agents. If the account proposed here is correct, the problem of causation doesn't force us to adopt any form of projectivism about causal relations. If

[8]Thus Price and Weslake write that any 'statistical asymmetry does a poor job of explaining why we don't (typically) deliberate with respect to past goals' (2009, 436).

PI is correct, causation is an objective phenomenon involving only certain patterns of objective correlations between events, so that the existence of non-perspectival objective causal facts is entirely compatible with the physics of our world.

Bibliography

Ahmed, A. (2010). Causation and decision. *Proceedings of the Aristotelian Society*, 110(2pt2):111–131.

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

Albert, D. (2013). Physics and chance. Unpublished manuscript.

Armstrong, D. M. (1983). *What Is a Law of Nature?* Cambridge University Press, Cambridge.

Armstrong, D. M. (1997). *A World of States of Affairs*. Cambridge University Press, Cambridge.

Boltzmann, L. (1897). Zu Hrn. Zermelo's Abhandlung ober die mechanische Erklärung irreversibler Vorgange. *Annalen der Physik*, 60:392–398. Reprinted in English translation in Brush (1966), pp. 238-245.

Bombelli, J., Lee, J., Meyer, D., and Sorkin, R. D. (1987). Spacetime as a causal set. *Physical Review Letters*, 59:521–524.

Brush, S. G., editor (1966). *Kinetic Theory*, volume 2. Pergamon Press, Oxford.

Carroll, J. (1991). Property-level causation? *Philosophical Studies*, 63(3):245–270.

Carroll, J. (1994). *Laws of Nature*. Cambridge University Press, Cambridge.

Carroll, J. (2009). Anti-reductionism. In Beebee, H., Hitchcock, C., and Menzies, P., editors, *Oxford Handbook of Causation*, pages 279–298. Oxford University Press, Oxford.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 13(4):419–437.

de Beauregard, O. C. (1977). Time symmetry and the Einstein paradox. *Il Nuovo Cimento*, 42B(1):41–64.

DeRose, K. (2010). The conditionals of deliberation. *Mind*, 119:1–42.

Diaconis, P. (1998). A place for philosophy? The rise of modeling in statistical science. *Quarterly of Applied Mathematics*, 56(4):797–805.

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

Dretske, F. (1977). Laws of nature. *Philosophy of Science*, 44:248–268.

Ducasse, C. J. (1926). On the nature and the observability of the causal relation. *Journal of Philosophy*, 23(3):57–68.

Dummett, M. (1954). Can an effect precede its cause? *Proceedings of the Aristotelian Society Supplement*, 28(3):27–44.

Dummett, M. (1964). Bringing about the past. *Philosophical Review*, 73:338–359.

Eagle, A. (2004). Twenty-one arguments against propensity analyses of probability. *Erkenntnis*, 60:371–416.

Eagle, A. (2007). Pragmatic causation. In Price, H. and Corry, R., editors, *Causation, Physics and the Constitution of Reality*, pages 156–190. Clarendon Press, Oxford.

Eagle, A. (2014). Generic causation. Unpublished manuscript. Available at http://dl.dropbox.com/u/6362052/generic-causation.pdf.

Earman, J. (1986). *A Primer on Determinism*. Reidel, Dordrecht.

Earman, J. (2008). Reassessing the prospects for a growing block model of the universe. *International Studies in the Philosophy of Science*, 22(2):135–164.

Eells, E. (1981). Causality, utility and decision. *Synthese*, 48(2):295–329.

Eells, E. (1982). *Rational Decision and Causality*. Cambridge University Press, Cambridge.

Eells, E. (1984). Metatickles and the dynamics of deliberation. *Theory and Decision*, 17(1):71–95.

Eells, E. (1991). *Probabilistic Causality*. Cambridge University Press, Cambridge.

Egan, A. (2007). Some counterexamples to causal decision theory. *Philosophical Review*, 116(1):93–114.

Elga, A. (2001). Statistical mechanics and the asymmetry of counterfactual dependence. *Philosophy of Science*, 68:313–324.

Elga, A. (2004). Infinitesimal chances and the laws of nature. *Australasian Journal of Philosophy*, 82(1):67–76.

Fales, E. (1990). *Causation and Universals*. Routledge, London.

Farr, M. and Reutlinger, A. (2013). A relic of a bygone age? Causation, time-symmetry and the directionality argument. *Erkenntnis*, 78:215–235.

Field, H. (2003). Causation in a physical world. In Zimmerman, D., editor, *Oxford Handbook of Metaphysics*, pages 435–460. Oxford University Press, Oxford.

Fine, K. (1975). Critical notice: *Counterfactuals*. *Mind*, 84:451–458.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, 71:5–19.

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

Frisch, M. (2007). Causation, counterfactuals, and entropy. In Price and Corry (2007), pages 351–395.

Frisch, M. (2010). Does a low-entropy constraint prevent us from influencing the past? In Ernst, G. and Hüttemann, A., editors, *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics*, pages 13–33. Cambridge University Press, Cambridge.

Gibbard, A. and Harper, W. (1978). Counterfactuals and two kinds of expected utility. In Hooker, C., Leach, J., and McClennen, E., editors, *Foundations and Applications of Decision Theory*, pages 125–162. Reidel, Dordrecht.

Glynn, L. (2009). *A Probabilistic Analysis of Causation*. PhD thesis, Oxford University.

Hájek, A. (1997). 'Mises redux' - redux: Fifteen arguments against finite frequentism. *Erkenntnis*, 45:209–227.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, 156:185–215.

Hájek, A. (2009). Fifteen arguments against hypothetical frequentism. *Erkenntnis*, 70(2):211–235.

Hall, N. (1994). Correcting the guide to objective chance. *Mind*, 103(412):505–518.

Hall, N. (2004). Two concepts of causation. In Collins, J., Hall, N., and Paul, L., editors, *Counterfactuals and Causation*, pages 225–276. MIT Press, Cambridge MA.

Hamilton, W. D. (1964). The genetical evolution of social behaviour I and II. *Journal of Theoretical Biology*, 7:1–32.

Hartle, J. (2005). The physics of now. *American Journal of Physics*, 73:101–109.

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

Hawking, S. and Ellis, G. (1973). *The Large Scale Structure of Space-Time*. Cambridge University Press, Cambridge.

Hesslow, G. (1976). Discussion: Two notes on the probabilistic approach to causality. *Philosophy of Science*, 43(2):290–292.

Hitchcock, C. (1996a). Causal decision theory and decision-theoretic causation. *Noûs*, 30(4):508–526.

Hitchcock, C. (1996b). The role of contrast in causal and explanatory claims. *Synthese*, 107:395–419.

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

Hitchcock, C. (2007). What Russell got right. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality*, pages 45–65. Clarendon Press, Oxford.

Hitchcock, C. (2012). Probabilistic causation. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*, Winter 2012 Edition. URL = http://plato.stanford.edu/entries/causation-probabilistic/.

Hitchcock, C. (2013). What is the 'cause' in causal decision theory? *Erkenntnis*, 78:129–146.

Hoefer, C. (2007). The third way on objective probability: A sceptic's guide to objective chance. *Mind*, 116(463).

Hoover, K. (2003). Nonstationary time series, cointegration, and the principle of the common cause. *British Journal for the Philosophy of Science*, 54:527–551.

Horwich, P. (1985). Decision theory in light of Newcomb's problem. *Philosophy of Science*, 52(3):431–450.

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

Ismael, J. (2008). Raid! Dissolving the big bad bug. *Noûs*, 42(2):292–307.

Ismael, J. (2009). Probability in deterministic physics. *Journal of Philosophy*, 106(2):89–108.

Ismael, J. (2011). A modest proposal about chance. *Journal of Philosophy*, 108(8):416–442.

Jeffrey, R. (1983). *The Logic of Decision*. University of Chicago Press, Chicago, 2nd edition. First edition published in 1965.

Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press, New York.

Joyce, J. (2012). Regret and instability in causal decision theory. *Synthese*, 187:123–145.

Keller, J. (1986). The probability of heads. *The American Mathematical Monthly*.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In Kitcher, P. and Salmon, W., editors, *Scientific Explanation*, pages 410–505. University of Minnesota Press, Minneapolis.

Kutach, D. (2002). The entropy theory of counterfactuals. *Philosophy of Science*, 69(1):82–104.

Kutach, D. (2007). The physical foundations of causation. In Price and Corry (2007), pages 327–350.

Kutach, D. (2013). *Causation and its Basis in Fundamental Physics*. Oxford University Press, Oxford.

Kvart, I. (2001). Causal relevance. In Woods, J. and Brown, B., editors, *New Studies in Exact Philosophy: Logic, Mathematics and Science*, volume II, pages 59–90. Hermes, London.

Lewis, D. (1973a). Causation. *Journal of Philosophy*, 70(17):556–567.

Lewis, D. (1973b). *Counterfactuals*. Blackwell, London.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13(4):455–476.

Lewis, D. (1980). A subjectivist's guide to objective chance. In Jeffrey, R., editor, *Studies in Inductive Logic and Probability*, volume 2, pages 263–293. University of California Press, Berkeley.

Lewis, D. (1981). Causal decision theory. *Australasian Journal of Philosophy*, 59:5–30.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61:343–377.

Lewis, D. (1986a). Postscripts to 'A Subjectivist's Guide to Objective Chance'. In *Philosophical Papers*, volume II, pages 114–132. Oxford University Press, Oxford.

Lewis, D. (1986b). Postscripts to 'Causation'. In *Philosophical Papers*, volume II, pages 172–213. Oxford University Press, Oxford.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103:473–490.

Lewis, D. K. (1986c). Events. In *Philosophical Papers*, volume II, pages 241–269. Oxford University Press, Oxford.

Loewer, B. (1996). Humean supervenience. *Philosophical Topics*, 24:101–127.

Loewer, B. (2004). David Lewis's Humean theory of objective chance. *Philosophy of Science*, 71(5):1115–1125.

Loewer, B. (2007). Counterfactuals and the second law. In Price and Corry (2007), pages 293–326.

Loewer, B. (2012). Two accounts of laws and time. *Philosophical Studies*, 160(1):115–137.

Maslen, C. (2004). Causes, contrasts and the nontransitivity of causation. In Collins, J., Hall, N., and Paul, L., editors, *Counterfactuals and Causation*, pages 341–357. MIT Press, Cambridge MA.

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

Meek, C. and Glymour, C. (1994). Conditioning and intervening. *British Journal for the Philosophy of Science*, 45(4):1001–1021.

Mellor, D. H. (1995). *The Facts of Causation*. Routledge, London.

Menzies, P. (1989). A unified account of the causal relata. *Australasian Journal of Philosophy*, 67:59–83.

Mill, J. S. (1916). *A System of Logic: Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*. Longman, Green and Co., London, 8th edition. First edition published in 1843.

Ney, A. (2009). Physical causation and difference-making. *British Journal for the Philosophy of Science*, 60:737–764.

Norton, J. (2007a). Causation as folk science. In Price and Corry (2007), pages 293–326.

Norton, J. (2007b). Do the causal principles of modern physics contradict causal anti-fundamentalism? In Machamer, P. and Wolters, G., editors, *Thinking about Causes: From Greek Philosophy to Modern Physics*, pages 222–234. University of Pittsburgh Press, Pittsburgh.

Norton, J. (2008). The dome: An unexpectedly simple failure of determinism. *Philosophy of Science*, 75:786–798.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In Rescher, N., editor, *Essays in Honor of Carl G. Hempel*, pages 114–146. Reidel, Dordrecht.

Papineau, D. (1985). Causal asymmetry. *British Journal for the Philosophy of Science*, 36(3):273–289.

Papineau, D. (1993). Can we reduce causal direction to probabilities? In Hull, D., Forbes, M., and Okruhlik, K., editors, *PSA: Proceedings of the Biennal Meeting of the Philosophy of Science Association 1992*, volume 2, pages 238–252, East Lansing, MI.

Papineau, D. (2001a). Evidentialism reconsidered. *Noûs*, 35(2):239–259.

Papineau, D. (2001b). Metaphysics over methodology - or, why infidelity provides no grounds to divorce causes from probabilities. In Galavotti, M. C., Suppes, P., and Costantini, D., editors, *Stochastic Causality*, pages 15–38. CSLI Stanford, Stanford.

Paul, L. A. (2000). Aspect causation. *Journal of Philosophy*, 97:223–234.

Paul, L. A. and Hall, N. (2013). *Causation: A User's Guide*. Oxford University Press, Oxford.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2nd edition. First edition published in 2000.

Pearl, J. and Verma, T. S. (1991). A theory of inferred causation. In Allen, J., Fikes, R., and Sandewell, E., editors, *Principles of Knowledge Representation and Reasoning*, pages 441–452, San Mateo CA. Morgan Kaufmann.

Popper, K. (1956). The arrow of time. *Nature*, 177:538.

Price, H. (1986). Against causal decision theory. *Synthese*, 67:195–212.

Price, H. (1991). Agency and probabilistic causality. *British Journal for the Philosophy of Science*, 42(2):157–176.

Price, H. (1992). Agency and causal asymmetry. *Mind*, 101(403):501–520.

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

Price, H. (2007). Causal perspectivalism. In Price, H. and Corry, R., editors, *Causation, Physics and the Constitution of Reality*, pages 250–292. Oxford University Press, Oxford.

Price, H. (2012). Causation, chance, and the rational significance of supernatural evidence. *Philosophical Review*, 121(4):483–538.

Price, H. (2014). Causation, intervention and agency - Woodward on Menzies and Price. In Beebee, H., Hitchcock, C., and Price, H., editors, *Making A Difference*. Oxford University Press, Oxford. Forthcoming.

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

Price, H. and Weslake, B. (2009). The time-asymmetry of causation. In Beebee, H., Hitchcock, C., and Menzies, P., editors, *Oxford Handbook of Causation*, pages 414–443. Oxford University Press, Oxford.

Putnam, H. (1975). Philosophy and our mental life. In *Mind, Language and Reality*, pages 291–303. Cambridge University Press, Cambridge.

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

Reid, D. D. (2001). Introduction to causal sets: An alternative view of spacetime structure. *Canadian Journal of Physics*, 79:1–16.

Russell, B. (1913). On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26.

Russell, B. (1948). *Human Knowledge*. Simon and Schuster, New York.

Salmon, W. (1980). Probabilistic causality. *Pacific Philosophical Quarterly*, 61:50–74.

Schaffer, J. (2000a). Causation by disconnection. *Philosophy of Science*, 67(2):285–300.

Schaffer, J. (2000b). Trumping preemption. *Journal of Philosophy*, 97(4):165–181.

Schaffer, J. (2001). Causation, influence and effluence. *Analysis*, 61(1):11–19.

Schaffer, J. (2005). Contrastive causation. *Philosophical Review*, 114(3):327–358.

Schaffer, J. (2009). Causation and laws of nature: Reductionism. In Sider, T., Hawthorne, J., and Zimmerman, D., editors, *Contemporary Debates in Metaphysics*, pages 82–107. Blackwell, Oxford.

Schaffer, J. (2014). The metaphysics of causation. In Zalta, E. N., editor, *Stanford Encyclopedia of Philosophy*, Summer 2014 Edition. URL = http://plato.stanford.edu/entries/causation-metaphysics/.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge MA.

Sober, E. (1985). Two concepts of cause. In Asquith, P. and Kitcher, P., editors, *PSA*, volume 2, pages 405–424, East Lansing. Philosophy of Science Association.

Sober, E. (1988). The principle of the common cause. In Fetzer, J., editor, *Probability and Causation: Essays in Honor of Wesley Salmon*, pages 211–228. Reidel, Dordrecht.

Sober, E. (2001). Venetian sea levels, British breads, and the principle of the common cause. *British Journal for the Philosophy of Science*, 52(2):311–346.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. MIT Press, Cambridge, MA, 2nd edition. First edition published in 1993.

Spohn, W. (2001). Bayesian nets are all there is to causal dependence. In Galavotti, M. C., Suppes, P., and Costantini, D., editors, *Stochastic Dependence and Causality*, pages 157–172. CSLI Publications, Stanford.

Stalnaker, R. (1981). Letter to David Lewis. In Harper, W., Stalnaker, R., and Pearce, G., editors, *Ifs: Conditionals, Belief, Decision, Chance and Time*, pages 151–152. Dordrecht.

Strevens, M. (2011). Probability out of determinism. In Beisbart, C. and Hartmann, S., editors, *Probabilities in Physics*, pages 339–364. Oxford University Press, Oxford.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North Holland Publishing Company, Amsterdam.

Tooley, M. (1977). The nature of laws. *Canadian Journal of Philosophy*, 7:667–698.

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

Tooley, M. (1990). Causation: Reductionism versus realism. *Philosophy and Phenomenological Research*, 50:215–236.

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