# A Deliberative Account of Causation:

# How the Evidence of Deliberating Agents Accounts for Causation and its Temporal Direction

Alison Sutton Fernandes

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

Columbia University

2016

# ABSTRACT

## A Deliberative Account of Causation:

## How the Evidence of Deliberating Agents Accounts for Causation and its Temporal Direction

Alison Fernandes

In my dissertation I develop and defend a deliberative account of causation: causal relations correspond to the evidential relations we use when we decide on one thing in order to achieve another. Tamsin's taking her umbrella is a *cause* of her staying dry, for example, if and only if her deciding to take her umbrella for the sake of staying dry is *adequate grounds for believing* she'll stay dry. I defend the account in the form of a biconditional that relates causal relations to evidential relations. This biconditional makes claims about causal relations, not just our causal concepts, and constrains metaphysical accounts of causation, including reductive ones. Surely we need science to investigate causal structure. But we can't justify any particular account of causation independently of its relevance for us. This deliberative account explains why we should care about causation, why we deliberate on the future and not the past, and even why causes come prior in time to their effects.

In chapter 1 I introduce the motivations for the project: to reconcile causation and our freedom as agents with the picture of the world presented by physics. Fundamental physics makes no mention of causes. And the lawlike character of the world seems to rule out freedom to decide. My dissertation offers a combined solution—I explain our freedom in epistemic terms and use this freedom to make sense of causation.

In chapter 2 I draw on philosophy of action and decision theory to develop an epistemic model of deliberation, one based in requirements on belief. If we're to deliberate, our beliefs can't epistemically settle how we'll decide, yet our decisions must epistemically settle what we'll do. This combination of belief and suspension of belief explains why we rationally take ourselves to be free to decide on different options in deliberation.

In chapter 3 I defend this model from near rivals that also explain freedom in terms of belief. Accounts of 'epistemic freedom' from David Velleman, James Joyce and Jenann Ismael appeal to our justification to form beliefs 'unconstrained' by evidence. Yet, I will argue, these accounts are susceptible to counterexamples and turn out to rely on a primitive ability to believe at will—one that makes the appeal to justification redundant. J. G. Fichte's Idealist account of freedom, based in a primitive activity of the 'I', nicely illustrates the kind of freedom these accounts rely on.

In chapter 4 I develop the epistemic model of deliberation into a deliberative account of causation. I argue that A is a type-level cause of B *if and only if* an agent deciding on a state of affairs of type A in 'proper deliberation', for the sake of a state of affairs of type B, would be good evidence of a state of affairs of type B obtaining. This biconditional explains why we should care about causal relations—they direct us to good decisions. But existing accounts of causation don't adequately explain why causation matters. James Woodward's interventionist account explicates 'control' and 'causation' in the very same terms—and so can't appeal to a relation between them to explain why we should care about causal relations. David Lewis' reductive account relies on standards for evaluating counterfactuals, but doesn't motivate them or explain *why* a causal relation analysed in these terms should matter.

Delivering the right verdicts is not enough. The deliberative account explains why causation matters, by relating causal relations to the evidential relations needed for deliberation.

In chapter 5 I use the deliberative account to explain causal asymmetry—why, contingently, causes come before their effects. Following an approach from Huw Price, because deliberation comes prior to decision, deliberation undermines evidential relations towards the past. So an agent's deciding for the sake of the past in proper deliberation won't be appropriate evidence of the past, and backwards causation is not implied. To explain why deliberation comes prior to decision, I appeal to an epistemic asymmetry, one that is explained by statistical-mechanical accounts of causation in non-causal terms. But statistical-mechanical accounts still need the deliberative account to justify why the relations they pick out as causal should matter to us.

The deliberative account of causation relates causal relations to the evidential relations of use to deliberating agents. It constrains metaphysical accounts, while revealing their underlying explanatory structure. And it does not rule out explanations of causal asymmetry based in physics, but complements them. Overall this project makes sense of causation by foregrounding its relevance for us.

# Contents

## Graphs and Illustrations

## Acknowledgments

Many people have helped and supported me throughout my graduate years in ways I can't even begin to adequately describe or thank. But let me acknowledge what I can. I'm especially grateful to my advisor, Achille Varzi, for his endless encouragement, clear guidance, good-humour and detailed comments on so many drafts. My ideas would have made very much less sense without his help, and he did this all while keeping in mind the very human side of philosophy. I'm also grateful to Jenann Ismael and John Collins for their inspiring conversations, incisive criticisms and enthusiasm throughout the project; David Albert for first introducing me to these topics and helping me see the deeply philosophical side of physics; and Huw Price and Wolfgang Mann for many years of careful guidance and support. I'm also grateful to a number of others who read chapters and early drafts of this work and went out of their way to provide helpful and nuanced comments and criticisms, including Barry Loewer, Patricia Kitcher, Matthew Simpson, Arif Ahmed, John Maier, Rush Stewart, Katja Vogt, Thimo Heisenberg, Lucy Campbell, Brian Hedden, Michael Hicks, Fred Neuhouser, Carol Rovane and Isaac Levi. I'd also like to thank several others for generously sharing their thinking and helping me throughout my graduate years, including Borhane Blili Hamelin, Shyane Siriwardena, Jorge Morales, Zac Al-Witri, Jonathan Fine, Georgie Statham, Brad Weslake, Matthias Frisch, Sidney Felder, Thomas Blanchard, Heather Demarest, David Braddon-Mitchell, Kristie Miller, Paul Redding, Kathryn Tabb, Philip Kitcher and Hugh Mellor. I'd like to thank the staff, students and faculty of the philosophy departments of Columbia University, the University of Cambridge and at the University of Sydney—the last two where I was welcomed as a visiting student. I'd also like to thank audiences at Columbia University, the University of Cambridge, Rutgers University, the University of Arizona and the University of Sydney, for comments, criticisms and helpful suggestions.

I received personal support and encouragement from many of the people mentioned above. But let me thank a few more people whose thoughts, love and care have been particularly important to me: Jacinta van den Berg, Grant Dempsey, their daughter Sunday, Emma Gatehouse, Usha Nathan, Gal Katz, Laura Martin, Kyle Mitchell, Ariadna Pop, Kathryn Rickus and my Mum and Dad.

Parts of the dissertation are currently forthcoming in print. Sections of chapter 3 will appear in 'Varieties of Epistemic Freedom', *Australasian Journal of Philosophy*. Sections of chapter 5 will appear in 'Time, Flies, and Why We Can't Influence the Past', in Barry Loewer, Brad Weslake and Eric Winsberg (eds.) *Time's Arrows and the Probability Structure of the World.* Cambridge, Mass.: Harvard University Press.

*To friends near and far*

# Chapter 1

# Introduction: Freedom and Causation

## 1.1 The Problem with Causation

Causation plays an essential role in our scientific and everyday understanding of the world. We discover causal relations, construct causal theories, and explain effects by appealing to their causes.[1] And yet, I will argue, we lack a good philosophical account of causation. Why? Because we lack an account that explains why causation *should* matter to us—and this is crucial for making sense of causation. In the following chapters, I'll develop and defend a deliberative account of causation—one that accounts for causation by considering its relevance for deliberation. I'll argue, in fact, that causal relations correspond to the evidential relations agents need when they decide on one thing in order to achieve another. Say Tamsin is deliberating on whether to take her umbrella in order to stay dry. I'll argue that Tamsin's taking her umbrella is a *cause* of her staying dry if and only if her deciding to take her umbrella (in an appropriate deliberation) would be *adequate grounds for believing* she'll stay dry. There is an important correspondence between the belief-justifying (what I'll call 'evidential') relations that hold for deliberators like Tamsin, and causal relations.

I'll defend this account in the form of a biconditional that relates causal relations to evidential relations. This biconditional makes claims about causal relations, not just causal concepts. It tells us how causal and evidential relations relate. While the biconditional can be used to reduce causal relations to evidential relations, that is not my main aim. The biconditional is useful and interesting because of what it explains—it explains why causal relations should matter to us. In the remainder of this introduction, I'll say more about the

[1] I'll use the term 'causation' for the general relation of causation. I'll use the term 'causal relations' for particular instances of causation. I won't use the term 'causality' at all.

motivations for this project (sections 1–2), before outlining my accounts of deliberation and causation (sections 3–4). I'll end (section 5) with brief summaries of the chapters to come.

Why do we need an account of causation, when we seem to have a good understanding of causation from science? We discover causal relations using experiments, whether in labs, studies or even everyday life—such as when a chemist tests to see if adding too much acid caused the wrong product to precipitate, or when statistical data are used to determine if a new drug causes a decrease in cholesterol levels, or even when you experiment with having less coffee to see if it improves your sleep. We also construct theories that describe and explain causal relations.[2] Biologists map the causal relations involved in photosynthesis, physicists use models to explain why absorbing a photon causes an electron to jump to a higher energy level, and I might hypothesise about how too little sunlight is affecting my plant's health. Talk of causation is ubiquitous. But why would we need an account of it?

The problem is that when we look to our best candidates for fundamental scientific theories, those that aim to be universal in scope and explain the success of other scientific theories, causes don't appear. Fundamental physical theories don't identify particular states as causes, and others as effects. What we have instead are dynamical equations that relate different states of affairs at different times—no mention of causes. Furthermore, when we look at how fundamental physical laws operate, they seem to have precisely the wrong features to pick out causes. As I'll explain, following Russell (1912–13), fundamental laws only determine states of affairs at other times, given information about the entire global state of

[2] We appeal to causes for a variety of purposes, including giving explanations. I will refer to both causes and causal explanations in what follows, taking it as a minimal condition that causal explanations correctly describe causal relations. But my main concern is with causal relations, not with causal explanation *per se*.

affairs. Yet causes seem to be local states of affairs that directly determine their effects. Secondly, fundamental physical laws, in relevant respects, work equally well in *both* temporal directions. Yet causes seem to come prior to their effects. So there's a problem understanding how causation fits into the world, given the picture presented by fundamental physics. This is why we need an account of causation. We need an account that makes sense of what function causation serves, beyond the predictions and derivations of fundamental physics.

Let's look in more detail at how physical laws operate. Imagine a closed system consisting of 26 billiard balls bouncing off one other. For simplicity, we'll ignore friction and electrodynamics, take the collisions to be elastic (without loss of kinetic energy), and take the system to be described by simple Newtonian laws of motion.[3] Say billiard ball *A* knocks into stationary billiard ball *B*, and then billiard ball *B* moves off. It seems that the movement of *A* caused the movement of *B*. But there is nothing in the fundamental physical laws to tell us this. What the fundamental physical laws tell us is that given the positions and velocities of all 26 billiard balls at one time, $t_1$, this is what their positions and velocities will be at another time, $t_2$. In other words, the dynamical equations relate states of affairs of the whole system. They do not select out particular events as causes and other events as effects. So when we ask a question such as, 'what caused ball *B* to move off at velocity *v* at time $t_2$?', what answer can we expect from the theory? None. Newtonian mechanics does not tell us what the causes are. The theory does tell us which of the balls collided with ball *B*. But this by itself doesn't tell us what caused *B*'s motion: some further analysis of causation is required. If ball

[3] While Newtonian mechanics is known to be false when taken as a theory of fundamental physics, it serves as a good illustration of points that apply even under more sophisticated candidates for fundamental theories. I return to this point below.

*A* collides with ball *B*, we may want to say that ball *A*'s motion caused ball *B*'s subsequent motion. And we may want to call ball *C*'s motion a cause of *B*'s, if ball *C* previously collided with *A*. But Newtonian mechanics gives us no direct guide to such choices. It says nothing explicitly about causes and effects.

That's the first problem: fundamental physical theories don't mention causes. Here's a second closely related problem. Fundamental physical theories, such as Newtonian mechanics, relate the whole state of a system at one time with its whole state at another time. They relate *global* states of affairs. They don't relate individual local states of affairs to one another. They don't tell us what ball *B*'s motion will be, given merely the motion of ball *A* and *B*. Let's consider our billiard balls again. It's tempting to say that, by the laws of Newtonian mechanics, if ball *A* moves in such and such a way, then stationary ball *B* waiting nearby *has to* move off at velocity *v*. But this is not the case. It matters to ball *B*'s subsequent motion that ball *D* is off to the side, failing to collide with it. And similarly for all the other 23 balls. Newtonian mechanics does not allow us to derive the motion of ball *A*, or *any other part* of the system, without knowing information about *all* the other balls at a given time.[4] If we want local causes to 'determine' or 'necessitate' their effects, we have a problem. Unless we are willing to take the entire state of the system at a given time to be the cause of its state at another time, and give up the project of individuating causes more finely, the kind of determinacy we find in fundamental physics is not the kind we seem to find in causation.

[4] More precisely, one needs a linear combination of information about the positions and velocities of all the other balls.

Bertrand Russell expressed precisely this concern in his influential 'On the Notion of Cause' (1912–13).[5] Russell argued that if causes had to necessitate their effects, nothing less than the full state of a system at a one time could be identified as a cause. But science investigates general types of events—those that can be repeated—not global states. So, he argued, there can't reasonably be a cause-based science:

> In order to be sure of the expected effect, we must know that there is nothing in the environment to interfere with it. But this means that the supposed cause is not, by itself, adequate to insure the effect. And as soon as we include the environment, the probability of repetition is diminished, until at last, when the whole environment is included, the probability of repetition becomes almost nil.
>
> (1912–13, pp. 7–8)

Either the causes would be so extensive and detailed as to be unique, and not the subject of scientific investigation, or causes would not necessitate their effects. Bracketing the question of whether causal science might still be possible, Russell gave us a powerful argument for why local individual causes cannot necessitate their effects, given the laws of fundamental physics.

Is it a peculiar feature of Newtonian mechanics that it is 'global' in this sense and does not individuate local causes? Might there be better candidates for fundamental physical theories that are more local? While one can certainly posit theories of this form, none of the theories we take to be anything like plausible candidates for fundamental physics at our world have this character. Instead, they retain the same global form. To see why, firstly note that the problem with globalism arises even before we get to an explicit formulation of the laws, such as Newtonian mechanics. The conflict arises whenever parts of a system can potentially interfere with relations between other parts. If a theory is to correctly predict what will

[5] For further discussion of Russell's arguments, see Earman (1976), Field (2003), Hitchcock (2007) and Eagle (2007).

happen next, using causes, and causes are to necessitate their effects, then all the potentially interfering parts have to be included as causes. In the billiard ball example above, I didn't need to appeal explicitly to laws like $F = ma$. I appealed instead to intuitions about our everyday experience with situations like these, where we know we have to consider potential influences from the whole system if we are to work out what happens next.

Secondly, it might seem like the *determinism* of Newtonian mechanics is playing a role—the fact that only one subsequent state of the system is compatible with its current state and the laws.[6] But determinism is not a crucial feature. Indeterministic theories such as GRW and other chancy versions of quantum mechanics have the same global structure. The fact that more than one subsequent state is compatible with a system's present state and the laws does not mean that individual local causes enter the picture. The best that such theories can do in deriving subsequent states still requires information concerning every part of the system. Even if we move to indeterministic laws, and accept that causes can't strictly 'necessitate' their effects, the problem remains.

Thirdly, these arguments don't rely on a commitment metaphysical necessity of the kind Hume argued against ([1738], book I part III). I haven't relied on any particular metaphysics of lawhood. One can even be Humean and deny that the laws metaphysically necessitate. The problem is about relating the kind of necessity found in the laws to the kind of necessity found in causation, however deflationarily we characterise each of these. The problem is that there is no clear relation between the two.

[6] This characterisation of determinism is rough—see Earman (1986) for a rich discussion of how to characterise determinism, and whether Newtonian mechanics is in fact deterministic.

Altogether, the problem of relating local causes to global laws is not due to particular features, formulations or a particular metaphysics of fundamental physical theories. It is due to deep features of what we expect a fundamental physical theory to do. What we ask of a fundamental physical theory is that given full information about the state of the world at a given time, there is no further structural information than is provided by the theory that would allow us to tell more about what happens at a different time.[7] Put loosely (and more metaphysically) as much necessity as there is in the system should be reflected in the laws. We will be limited by the kinds of structures and patterns found in nature—there may be irreducibly chancy laws. But a fundamental theory aims to derive as much as possible about events at one time, given information about other times and sufficiently general laws. And it is no surprise that doing so accurately requires a vast amount of detailed information about what is happening at other times. We need fundamental physical theories to be global.

Might a fundamental physical theory require less than information about each part of the system to determine what happens at another time? Could we then zero in on individual causes? The prospects don't look promising. We might imagine a toy theory that requires less than information about the full system to derive subsequent states—such as Newtonian mechanics with an additional restriction on faster-than-light influence. With such a theory, we would not need information about the full system to determine subsequent local states. The backwards or forwards light-cone of the local state would suffice, since there would be insufficient time for any state beyond the light-cone to influence it.[8] But this still leaves an

[7] It is of course no easy task to say what counts as such structural information and whether it can be characterised, for example, in terms of generality and simplicity as a Humean analysis of laws suggest. I take no stand here on the nature of laws themselves, noting that various analyses of laws are compatible with them playing the kind of role I appeal to here.

[8] More precisely, we would need information about each point in the light cone—see Field (2003).

ever-expanding slice of the world as a potential influence, and so a potential cause, a slice that increases in radius at the speed of light. The set of potential causes is still too large for the theory to pick out individual causes.

Nor is a restriction to local states justified by current physics. General relativity doesn't involve a restriction on faster than light travel or influence (Maudlin 2004, ch. 3). While ordinary particles cannot be accelerated faster than the speed of light, the theory does allow for particles that travel faster, as well as for faster than light influence and information transfer. Furthermore, any theory that is able to reproduce the experimental results of Bell inequalities, predicted by quantum mechanics, must include correlations at space-like separations (separations that light cannot traverse). If these correlations are to be accounted for by a fundamental physical theory, the theory must require more than local information for its application.

In fact, when we move from Newtonian mechanics to more plausible candidates for fundamental physical theories, the situation becomes, if anything, much worse. Quantum mechanics, in its various interpretations and variations, posits entangled states, implying that the full state of a system cannot even be described in terms of the states of its individual local parts. Even before we get as far as prediction and explanation, if anything like quantum mechanics is right, fundamental science is not based on local individual states of affairs at all. While the hope is that any results about causation we get from the Newtonian case will extend to quantum mechanics and general relativity, we will not be missing out on a straightforward solution to the problem of causation by beginning with Newtonian mechanics.

Finally, there's a third problem relating causation to fundamental physics—reconciling the temporal asymmetry of causation with the temporal symmetry of fundamental laws. An important feature of Newtonian laws is that they are temporally reversible. Newtonian laws do not distinguish between a forward and a backwards direction in time—they take the same form in either direction. For example, if one were to film a movie of our idealised billiard balls bouncing around, and then play the movie backward, their motion would still be perfectly described by the same Newtonian laws. As Russell noted, 'the future "determines" the past in exactly the same sense in which the past "determines" the future' (1912–1913, p. 15). Yet we take causes to come prior in time to their effects. Fundamental physics does not seem to have the characteristic temporal asymmetry that we find with causation.

There are candidates for fundamental physical theories that do employ time-*irreversible* laws. Variations on quantum mechanics, such as GRW, employ stochastic chancy laws that apply in one temporal direction only. But, as I'll argue in chapter 5, these asymmetries aren't well-placed to explain the pervasive macroscopic temporal asymmetry of causation. Moreover, even if the asymmetry of laws could directly explain the asymmetry of causation, the problem with globalism would remain. Even if a theory explicitly distinguished between later and earlier states, and we stipulated that later states of the system can't be causes of earlier states, we would still be left with whole previous time slices causing later ones. We wouldn't be able to pinpoint individual causes using the theory.

Russell used similar points to argue for the wholesale elimination of causal talk from science and philosophy. If causes aren't found in fundamental physics, talk of them should be expunged altogether—it does more harm than good (1912–13, p. 1). But we needn't go so far. Causation may still be useful in science, even if fundamental theories do not employ causal terms, and even if causation does not imply necessitation by physical laws. Nor do the above arguments show that fundamental physics is somehow incompatible with causation, such that there is a contradiction between causal claims and the claims of fundamental physics, or an incompatibility between the kind of determinacy present in causes and the laws. In fact, a sensible desideratum on a philosophical account of causation is that it guarantees that causation is compatible with fundamental physical laws (or at least our best guess as to the form these will take).

But compatibility does not mean that we can look to our fundamental theories alone to distinguish causes, tell us what makes one thing the cause of another, or tell us how causation and fundamental laws relate. Russell's arguments set up an important challenge: given that causes do not feature in fundamental science, and do not carry the kind of necessity present in fundamental laws, how do they feature in our scientific picture of the world?

### 1.2 The Problem with Agency

There is a second concern that motivates this project. How do we make sense of free agency, particularly our apparent freedom in deliberation, given the picture presented to us by physics? As we deliberate about whether to do this or that, different possibilities appear to us as options we are free to decide on. As you read this page, I imagine, it seems to you that you can choose whether to continue reading or take a break. You appear free to make such choices and determine how the world subsequently goes. What might account for this apparent freedom? One thought is that the world really is radically open. We seem free because the world itself, like us, is ‘undecided’ about which way to go; there is nothing in its current state that determines whether it goes one way or another. Until we decide, different alternative futures are compatible with its current state. The indeterminacy of the world is then reflected in our experience as agents—what happens next seems ‘up to us’ because our decisions are required, as additional features beyond the world’s present state, to determine how the world goes.

But a problem immediately arises, concerning the character of physical laws. As we’ve seen, Newtonian laws relate the entire state of a system at one time to its state at any other time. Given the state of the universe at any one time, and the laws, there is nothing further required to determine everything that happens in the universe at any other time.[9] This means that facts about what happened long before you were born, combined with the laws, are enough to determine all your subsequent decisions and actions. There is nothing undecided about which way the world is to go.[10] Your experience of freedom does not reflect any actual freedom you have.

Various alternatives have been explored to explain how deterministic laws might be compatible with our freedom as agents. For example, our freedom as agents might be closely connected to moral responsibility and the appropriateness of praise or blame, which the determinacy of physical laws seems to have little bearing on (Strawson 1974). Or our freedom might be connected to a lack of coercion, and to whether our actions are

[9] Assuming deterministic laws for the moment and noting exceptions explored in Earman (1986).

[10] van Inwagen presents a similar line of thought in his ‘consequence argument’ (1983). If you have no control over the laws or the state of the world before you were born, and the state of the world before you were born and the laws determine the state of the world at all other times, then you have no control over the state of the world at other times.

appropriately caused by our desires and decisions (Hume [1748], sec. 8; Frankfurt 1971). Or freedom might be connected to guidance by a reasons-responsive mechanism (Fischer 2006), or to our ability to reflectively endorse our desires (Korsgaard 1996). Some of these 'compatibilist' alternatives have become so familiar, that it might seem naïve to think there is even a *prima facie* conflict between our freedom and the kind of determinism present in the laws. So it is worth pausing on just how compelling the above conception of freedom appears. There seems something right in the idea that it is up to us, as an additional fact about the world beyond its previous state and laws, how we decide, and so how the future of the world is to go—something reflected in an experience as we deliberate. This appearance of freedom needs explaining.

Moreover, these compatibilist solutions don't straightforwardly explain our experience of freedom as we deliberate. Compatibilist responses tend to highlight the moral dimension of practical reasoning. For example, Strawson attempts to show that metaphysical concerns over the nature of free will do not endanger our moral practices. Other accounts aim at giving more or less precise specifications of the conditions under which we are appropriately held morally responsible. In the hands of these compatibilists, the debate concerning free will has shifted away from metaphysical concerns over determinism, and towards normative and ethical concerns about moral responsibility. Indeed, many compatibilists speak of the freedom they're concerned with as simply that which is required for moral responsibility (Strawson 1962; Frankfurt 1971; Korsgaard 1996; Fischer 2006). But explaining why you can be appropriately held responsible is not the same as explaining why it seems to you that you can decide and act in different ways. While one may be able to build an explanation of how things seem for the deliberating agent from attributions of moral responsibility, the steps required are not trivial. Alternatives are worth exploring.

There are also problems with other straightforward attempts to explain our apparent freedom in deliberation. Firstly, it might seem that only *deterministic* physical laws are incompatible with free agency—because only deterministic laws completely constrain the state of the world at all other times. If fundamental physical laws are indeterministic, as quantum mechanics suggests, then several different futures are compatible with the current and previous states of the world.[11] So perhaps our apparent freedom can be explained by the indeterminism of the laws?

However, appealing to indeterministic quantum mechanical laws does not explain our apparent freedom. Firstly, the indeterminism in quantum mechanics is largely hidden from investigations that proceed at ordinary macroscopic energy and length scales—explaining why quantum phenomena took so long to discover.[12] Any fundamental physical theory must explain why deterministic Newtonian laws work so well, and so cannot allow for large-scale deviations from deterministic behaviour at the kind of energy and length scales at which agents typically operate. The environment and stimuli we respond to are readily characterisable in macroscopic ways. And so are our behavioural outputs, in the form of bodily movements, as well as much of our inner workings. While quantum-mechanical phenomena may play a role at the neurological level, we would need a positive reason to

[11] I'll bracket the issue of whether quantum mechanical laws are in fact indeterministic. Two prominent accounts of quantum mechanics employ deterministic laws, at least at the fundamental level—Bohmian mechanics and Everettian (Many Worlds) quantum mechanics.

[12] This is not to say that quantum effects don't turn up at the macroscopic level, or that we can't bring about correlations between macroscopic states and quantum states. This is precisely what happens in quantum-mechanical experiments. But the indeterminacy present in quantum-mechanical phenomena are still not readily apparent from macroscopic behaviour.

think that quantum indeterminism here would scale up to produce indeterminism at the macroscopic level. Absent such grounds, we have no reason to think that deterministic macroscopic laws would often be violated. So the conflict between laws and our apparent freedom as agents remains.

Secondly, even if there were macroscopic indeterministic laws, this wouldn't capture what we seem to require by free choice. For example, if the indeterministic laws imply that with 50% probability you choose chocolate ice cream, and with 50% probability you choose vanilla, in what sense is it up to you what ice cream flavour you choose? You can't always choose chocolate in similar set ups without diverging from the laws.[13] Provided we hold the laws fixed, our behaviour is still constrained in ways that seem incompatible with free agency.[14]

Here's another attempt to explain our apparent freedom as agents. Russell, as we saw, argued that we should jettison the concept of causation altogether (1912–13). He distinguished causation from fundamental physical laws, accepted only the latter, and was optimistic that by separating the two, the question of freedom and determinism could be easily resolved. Once one realised that laws were not the kind of things that pushed and pulled us around, forcing and compelling us as our typical experience of causes suggests, determination by laws would no longer seem even *prima facie* incompatible with freedom. But distinguishing between laws and causation is not enough to resolve the problem. Even if laws are not the kind of things that push and pull us around, they are still enough to constrain what we do.

[13] In epistemic terms, this is a case of disconfirming a probabilistic law.

[14] For further discussion of why quantum mechanics does not explain free agency, see Dennett (1984, pp. 135-6) and van Inwagen (2008, p. 263 ff.).

They imply that the world is not radically open to choice in deliberation, as our experience suggests.

Finally, it might seem that the conflict over freedom and the laws is due to a particular conception of lawhood—that physical laws 'govern' how the world evolves (Armstrong 1983; Tooley 1977). So it might seem we can avoid the conflict by adopting a less metaphysically robust Humean conception of lawhood and taking laws to merely describe the evolution of the world (Lewis 1973; 1983; 1994). If laws merely describe, and do not govern, perhaps this removes the tensions with freedom. But this appearance is misleading. Those who adopt descriptive laws still aim to explain our practices of holding laws fixed when we evaluate counterfactuals or work out the future evolution of a system. And this is what produces the tension with free choice. With the laws fixed, nothing more than the current state of the world is required to determine the evolution of the system, with as much as determinacy as is present in nature. Perhaps one of the above responses can ultimately explain our freedom or apparent freedom in deliberation, in a way compatible with physical laws. But there are enough concerns with these responses to make other explanations of our apparent freedom worth exploring.

### 1.3 What is an Epistemic Model of Deliberation?

My accounts of deliberation and causation are addressed to two problems: understanding how deliberative freedom and causation fit into the world, given the picture presented to us by fundamental physics. I offer a combined solution: I account for our apparent freedom in deliberation in epistemic terms. And I use this apparent freedom to make sense of causation.

By understanding how the world leaves room for apparent deliberative freedom, we can understand how it leaves room for causation.

I begin, in chapter 2, by developing and defending an 'epistemic model' of deliberation, which explains our apparent freedom in deliberation. This model characterises deliberation by placing constraints on what beliefs we can and can't reasonably have while deliberating. For example, I'll argue that we can't reasonably deliberate about what we're going to do if we're already certain of what we're going to do. Yet we must take our decisions to settle what we'll in fact do. This account is a 'model' in the sense that it doesn't attempt to fully characterise deliberation. Like models in science and architecture (rather than logic), it selectively highlights certain features, while neglecting others. In this case, the model highlights epistemic features of deliberation, while neglecting causal and normative features. By neglecting causal features, the model holds out the possibility for non-circularly accounting for causation in terms of deliberation. By neglecting normative features, the model does commit to a particular view of ethics or practical reasons for action. Even though my epistemic account is only a model, I'll argue that it does important work: it explains why the world appears open to choice in deliberation, and so accounts for the kind of apparent freedom required to make sense of causation.

My main interest is in exploring deliberative freedom in order to account for causation. For this reason, there are aspects of freedom and deliberation I won't explore. I won't take a strong stance on whether we *are* free in deliberation or not. This would require settling what such freedom should amount to—a weighty issue indeed. My claim is that the epistemic model can explain why we appear free, leaving open the question of whether this appearance is veridical. For the record, my view is that we should adopt a relatively deflated picture of what our freedom in deliberation amounts to, so the epistemic model *can* explain why we're free. But I won't have room to defend this view, or to argue against more robust conceptions of freedom—and this isn't my main purpose. My main purpose is to develop a sufficiently rich model of deliberation that proceeds in non-causal terms, and so can be used to give a non-circular deliberative account of causation.

Others have also appealed to ignorance of some form to explain our freedom in deliberation, our apparent freedom, the apparent indeterminateness of the future, or to give conditions on deliberation—including Baruch Spinoza ([1677], IIIp2s), Bertrand Russell ([1910], pp. 29–33; 1912–13, pp. 20–21; [1914], pp. 237–240), Ludwig Wittgenstein ([1921], 5.1362), Michael Dummett (1954), Isaac Levi (1986, ch. 4; 1997, chs. 2 and 4; 2000) and Tomis Kapitan (1986; 1989). But, to my knowledge, no one has previously used an ignorance-based account of deliberation to give an account of causation. This point is not obvious, however. In chapter 3, I'll do further work to distinguish my ignorance-based account from other accounts of deliberation that also appeal to beliefs, including accounts by David E. Velleman (1989a; 1989b), James M. Joyce (2002; 2007), Jenann Ismael (2002; 2007) and Huw Price (2012). These accounts explain our (apparent) freedom in deliberation by our ability to form beliefs about what we'll decide or do, unconstrained by evidence. We are unconstrained by evidence because the beliefs concerned are 'self-fulfilling'—they bring about their own truth. And because these beliefs constitute our decisions, we have 'epistemic freedom' over our decisions. This kind of freedom *has* been appealed to in giving agent-based accounts of causation (Ismael 2007; Price 2012). I'll reject these accounts of deliberation. I'll argue that they are distinct from ignorance-based accounts (*contra* Price

2012), that they are susceptible to counterexamples, and that they rely problematically on primitive or desire-based freedom. If we're to explain our apparent freedom in deliberation by appealing to beliefs, we should appeal instead to ignorance—not to the idea that we can form beliefs unconstrained by evidence.

The epistemic model of deliberation I develop is relevant for understanding causation, but it also has important implications for debates about deliberation and freedom in decision theory, philosophy of action and meta-ethics. The epistemic model constrains idealised models of deliberation—defending it means arguing against competing models, such as Joyce's (2002; 2007). Defending the model also means arguing against competing explanations of our apparent freedom to decide from philosophy of action, such as Velleman's (1989a; 1989b), as well as accounts from meta-ethics that appeal to normative beliefs, such as Korsgaard's (1996). Insofar as an epistemic model explains our apparent freedom in deliberation, normative beliefs aren't needed to. The epistemic model therefore undercuts arguments from our apparent freedom in deliberation to the existence of normative facts.

### 1.4 What is a Deliberative Account of Causation?

In section 1, I argued that there was a problem understanding how causation fits into the world, given the picture presented to us by fundamental physics. So we need an account of causation. But why would we need a specifically *deliberative* account of causation? In chapters 4 and 5, I'll argue that an account of causation that relates causation to deliberation satisfies important desiderata on accounts of causation. Firstly, a deliberative account can explain why causal relations should *matter* to us. Given that causes don't feature in the predictions and derivations characteristic of fundamental physics, and aren't needed to determine events, we need an account that makes sense of what other role causes play. According to the biconditional I defend, Tamsin's taking her umbrella is a cause of her staying dry, if and only if her deciding to take her umbrella (in proper deliberation) is good evidence of her staying dry—where good evidence is the kind of state that justifies belief. According to this biconditional, if Tamsin knows about the causal structure of the situation, she can use this knowledge to make decisions that are evidence for outcomes she seeks. This is why agents should care about causation—causal knowledge is needed for good decision-making.

In chapter 4, I'll argue that prominent existing accounts of causation don't adequately explain why causal relations should matter to us. James Woodward (2003) and Judea Pearl (2000) defend 'interventionist' accounts of causation. They analyse causal relations in terms of counterfactuals concerning what would happen to a variable under interventions. These accounts are non-reductive: they don't reduce causal relations to non-causal relations. My concern with these accounts is not that they are non-reductive. My concern is that they don't adequately explain why agents should care about causation. As I'll argue, they rely either on trivial definitional explanations, or appeal worryingly to primitive notions of agency.

Reductive accounts also fail to explain why we should care about causation. These accounts attempt to reduce causal relations to non-causal relations, typically fundamental laws and contingent features. David Lewis (1973; 1979), for example, reduces causal relations to counterfactuals, and then evaluates these counterfactuals using a 'comparative similarity ordering' between possible worlds. While Lewis takes the similarity ordering to be a metaphysical primitive, he identifies standards by which similarity is to be evaluated. The problem, I'll argue, is that Lewis fails to justify these standards. He uses apparently arbitrary criteria that leave it mysterious why a causal relation formulated in these terms should matter to us—even if it delivers the right verdicts. Other reductive accounts inherit this problem, including the more sophisticated statistical-mechanical accounts of David Albert (2000) and Barry Loewer (2007).

The deliberative account aims to do better. It explains why causation should matter to us, by relating causal relations to the evidential relations needed for deliberation. When we're deliberating, we should make decisions that are evidence for outcomes seek. When Tamsin is deciding whether to take her umbrella, for the sake of staying dry, she should decide to take her umbrella if and only if her decision to take her umbrella is in fact good evidence of her staying dry. If she then decides to take her umbrella, she will have made an appropriate decision. Similarly, when Suzy is deliberating about whether to throw her rock for the sake of breaking the bottle, she'll have made an appropriate decision to throw if and only if her decision to throw is good evidence of the bottle breaking. And when politicians consider whether to introduce carbon credits to reduce the rate of global warming, their decision will be appropriate if only if their decision to introduce carbon credits is in fact evidence of a reduction in the rate of global warming. Causal relations matter to us because knowledge of them directs us to good decisions.

The deliberative account has other advantages as well. It explains why causes come prior in time to their effects, for macroscopic causes at our world. This is another explanatory challenge that Russell's arguments set up—given fundamental laws work equally in both temporal directions, why do causes only operate in one direction? The deliberative account has an answer. In chapter 5, I'll argue that because deliberation comes prior to decision, deliberation breaks or 'undermines' evidential relations towards the past.[15] Say Tamsin typically takes her umbrella after there are storm clouds in the sky, so her decision to take her umbrella is usually good evidence of earlier storm clouds. Would a deliberative account imply (incorrectly) that Tamsin's taking her umbrella is a *cause* of the earlier storm clouds? I'll argue that if Tamsin is properly deliberating, her decision to take her umbrella for the sake of there being earlier storm clouds *won't* be evidence of earlier storm clouds. Why? Because proper deliberation requires that Tamsin doesn't have evidence of there already being storm clouds as she deliberates. And if she doesn't have this evidence there's no reason to think her decision now will be correlated with there being earlier storm clouds. So these ordinary cases don't imply backwards causation. By generalising Tamsin's case, I'll argue that the deliberative account explains why causes come prior in time to their effects at our world.

Finally, the biconditional I defend also acts as an important constraint on metaphysical accounts of causation—the deliberative account can function as a 'meta-account' of causation. The biconditional determines how causal relations relate to evidential relations. Provided one can show that this correspondence is satisfied for a given account of causation and evidence, we have an explanation for why causation, analysed in these terms, should matter to agents. In this sense, the deliberative account is compatible with first-order metaphysical accounts. But the deliberative account is still what does the work in explaining why causal relations, understood in these particular terms, should matter—and even why causes come before their effects. I'll argue in chapter 5, for example, that while a deliberative account is compatible with statistical-mechanical accounts of causation, the success of these

[15] This undermining relates to the 'screening-off' appealed to by Huw Price (1992) and other evidential decision theorists. But there are crucial differences, explored chapter 5.

accounts should be understood in evidential and agent-based terms. It is insofar as these accounts pick up on a correspondence between evidential and causal relations that they deliver plausible accounts of causation.

On the other hand, it also turns out that the deliberative account should appeal to resources from statistical-mechanical accounts. Statistical-mechanical accounts explain an epistemic asymmetry in terms of entropy. Under the deliberative account, this same epistemic asymmetry can be used to explain why we deliberate before we decide—and so why correlations towards the past are undermined by deliberation. In common with statistical-mechanical accounts, I'll argue that causal asymmetry is ultimately due to an initial entropic boundary condition on the universe. For this reason, statistical-mechanical accounts and deliberative accounts turn out to be a particularly neat package. Statistical-mechanical accounts relate causation, evidence and entropy. The deliberative account explains why these evidence-based causal relations matter—they matter to deliberating agents.

I won't defend a particular account of what evidential relations are (although I'll consider some options in chapter 4). So I won't advocate a particular account of what causal relations fundamentally are or reduce to. The deliberative account can be used as a half-way step to ultimately reducing causal relations to fundamental laws and contingent features, via evidential relations. But such a reduction is not mandatory. The deliberative account is compatible with a variety of reductive accounts, as well as with accounts in which causation is *irreducible*.

Why don't I plump for a particular metaphysics of causation? To lay my metaphysical cards on the table, I believe the important work of metaphysics is to relate things to one another, in a way that is illuminating—not to reduce one set of things down to other more fundamental things. I'm suspicious of the asymmetrical notions of reduction and fundamentality involved. But this doesn't mean I reject metaphysics entirely. My own deliberative account of causation aims to relate causal and evidential relations in a way that is illuminating and explanatory. Reductive metaphysical accounts more generally can be read as doing similar work. Humean accounts of laws, for example, can be read as relating contingent features to laws in a way that is explanatory—but without taking contingent features to be more fundamental. For this reason, I don't think causation can be ultimately reduced in the way many metaphysicians intend. But I remain hopeful that interesting work can be done relating causation to laws and contingent features, via the deliberative account.

The metaphysical aspirations of my account are one of the ways in which it differs from other agent-based accounts. The account I offer shares important features with accounts from Huw Price (1991; 1992; 1993; (Menzies and Price) 1993; 1996; 2007; (Price and Weslake) 2009; 2012; forthcoming) and Jenann Ismael (2007), as well as early manipulationist accounts from R. G. Collingwood (1940, pp. 296–312), Douglas Gasking (1955) and G. H. von Wright (1971, ch. 2). These all share the idea that we must make sense of causation through its relevance for agency, deliberation and control. Price, whose account is furthest developed, defends a 'perspectivalist' account, arguing that for causation to appear in our worldview, we must take up the perspective of an agent, capable of intervening in the world. But while Price and I agree on the relevance of agents for understanding causation, we disagree about the metaphysical upshots of an agent-based account. Price most often takes his agency theory to concern the function of causal talk, the assertibility conditions on the concept causation, or the genealogy of the concept. My own account is directly concerned with causation itself. It aims to deliver causal relations that appear from outside the perspective of the agent. And it aims to be compatible with metaphysical accounts of causation.

Overall, my deliberative account sits between various existing counterfactual accounts of causation, sharing features in common with each. In common with reductive metaphysical accounts, I am interested in causal relations themselves, and how causation relates to fundamental laws. But unlike reductive accounts, I don't think a reduction of causation to fundamental laws and contingent features gives us a sufficiently illuminating account of causation. We need to think about why causal relations matter. In common with interventionist account, I think causal relations should be thought of in terms of their relevance for manipulation and control. But unlike interventionist accounts, I don't think merely relating causal relations to one another is enough. In common with Price's perspectivalist account, I think we can't make sense of causation without thinking about its relevance for deliberation. But unlike Price, I don't think making sense of causation by appeal to deliberation means giving up all aspects of the metaphysical project.

### 1.5 Summary of Chapters

In the chapters to come, I present and defend an epistemic model of deliberation (chapters 2–3), and a deliberative account of causation (chapters 4–5). In chapter 2, I put forward an epistemic model of deliberation. This model explains our apparent freedom in deliberation using requirements on the beliefs we can and can't have while deliberating. I argue, for example, that Tamsin can't reasonably deliberate on taking her umbrella if she's already certain she won't. Her ignorance of what she'll do, along with other epistemic conditions, explains her apparent freedom to deliberate. The model appeals only to epistemic features, not metaphysical notions of freedom. And so it explains our experience of free deliberation in a way compatible with physical laws.

In chapter 3, I defend this model from rivals that also appeal to beliefs to explain our apparent freedom. These rival accounts appeal to the idea that agents can properly form certain beliefs 'unconstrained' by evidence. These beliefs constitute their decisions, giving agents a kind of 'epistemic freedom' over their decisions. This epistemic freedom is supposed to explain their freedom to decide. I'll argue that these accounts incorrectly imply agents have epistemic freedom even over beliefs formed on the basis of evidence, and rely on a faulty notion of justification. Underlying these troubles, I'll argue that these accounts rely problematically on a primitive ability to believe in different ways—they don't actually explain our freedom in purely epistemic terms.

In chapter 4, I develop my preferred epistemic model of deliberation into a 'deliberative account' of causation. Say Tamsin is deliberating about whether to take her umbrella, and properly considers her evidence. I argue that her taking her umbrella counts as a *cause* of her staying dry just in case her deciding to take her umbrella is good evidence she'll stay dry. This account explains why causal relations matter—causal relations are needed for good decision-making. But the account does not make causal relations merely subjective or up to us. Causal relations are as objective as the evidential relations we use when we reason, and are related to fundamental laws precisely because laws are also required for good evidential reasoning.

Finally, in chapter 5, I argue that the deliberative account can also explain why causes come *before* their effects. I begin by showing that if Tamsin is properly deliberating, her decisions won't be evidence for past outcomes she seeks. For example, Tamsin's decision to take her umbrella won't be evidence of earlier rain, if she decides to take her umbrella for the sake of there being earlier rain. Her deliberation, if it is proper, undermines the evidential relation that would otherwise obtain to the past. So a causal relation that corresponds to the evidential relations that hold for deliberators like Tamsin won't run backwards.

Ultimately, my project aims to make sense of our apparent freedom in deliberation and use this apparent freedom to make sense of causation. So we can understand how causation and our apparent freedom fit into the world, given the picture presented to us by science.

# Chapter 2

## An Epistemic Model of Practical Deliberation

*Experience teaches us no less clearly than reason, that men believe themselves to be free, simply because they are conscious of their actions, and unconscious of the causes whereby those actions are determined.*

Baruch de Spinoza ([1677], IIIp2s)

In this chapter, I present and defend an epistemic model of practical deliberation. This model characterises deliberation by placing constraints on the beliefs an agent can and can't reasonably have while deliberating. I'll argue, for example, that agents can't reasonably deliberate on what to do when they're already certain of what they're going to do. Yet they must take their decisions to epistemically settle what they'll do. This model does not attempt to capture all aspects of deliberation. But it captures enough, I will argue, to explain why different options appear available to be decided on in deliberation, and so why the world appears open to choice. This is precisely the kind of apparent freedom we need to make sense of causation, and to understand how our apparent freedom in deliberation fits into a scientific picture of the world.

The chapter proceeds as follows. In section 1, I'll introduce the kind of deliberation I'm concerned with and the general approach I'll take. In section 2, I argue for a set of necessary epistemic conditions on deliberation, which constitutes the epistemic model. In section 3, I show that the model is compatible with practical norms playing a role in deliberation. In section 4, I show how the model explains our apparent freedom in deliberation. And finally, in section 5, I advocate a certain picture about how practical and theoretical reason relate: practical deliberation is only allowed for by pre-existing gaps in our (theoretical) beliefs about the world.

### 2.1 What is Practical Deliberation?

In this chapter, I present a model of practical deliberation. I take the practical deliberation I'm concerned with to have the following features. It is a decision-making process in which an agent deliberates between different options that she takes to be available, and which she appears free to decide on. For example, Kyle might deliberate over whether to be in London or Edinburgh next week. Shyane might deliberate about whether to have milk in the fridge or not. In each case, possibilities appear to the agent as available options he or she is free to decide on. These options concern what is *to be* or *to be made* the case, where this is usually expressed in the future tense form of what one will do or what will be the case.[16] Practical deliberation of this form relates closely to action—to our attempts to make the world fit with our decisions—hence, it is *practical.* There are other activities that may rightly be called practical deliberation, but which I won't consider here—these include considering one's ends or desires and working out what options are available. I take deliberation to be a process that occurs over time, which typically conclude in decision.[17] Deliberation may be interrupted and returned to. A decision can be given up and deliberation restart. The decision may be implemented either at the end of deliberation, at some later point, or over a period of time. Decisions can be snap decisions, require short or long periods of deliberation, or be part of long term planning.

As I go on to develop the epistemic model of practical deliberation, I'll appeal primarily to work in decision theory rather than philosophy of action. This is for a number of reasons.

[16] While we often express practical decisions as concerning what 'will be', I will avoid such future-tensed language as far as possible. I allow for the possibility that we can decide on past state of affairs in order to give a non-trivial explanation of why we deliberate only on the future (chapter 5).

[17] But I don't assume that the conclusion of deliberation comes temporally *after* deliberation—the fact that decision comes after deliberation is to be explained (see chapter 5).

Firstly, while philosophers of action typically assume a causal structure to the world that agents have ready access to, there are decision theorists who do not take this structure for granted—particularly evidential decision theorists. If we are interested in accounting for causation in terms of deliberation, we would do well to assume as little causal structure as possible to begin with. Secondly, the minimal conditions used by decision theorists are compatible with a variety of approaches to ethics and normative reasons for action. They do not settle questions such as whether reason-giving relations are objective features of the world, or based on our desires. I am interested in specifying conditions on deliberation that theorists with different views of normativity can accept, and so the very minimal conditions used in decision theory are preferable to the more normatively-laden views developed in philosophy of action. While decision theorists are still concerned with principles of rationality that are taken to be normative, an attempt is made to be otherwise neutral concerning reasons for action.

Because I bracket causation and practical normativity as much as possible, the account of deliberation I develop is going to look strange and perhaps misguided to those more familiar with accounts of deliberation from ethics, meta-ethics and philosophy of action. I encourage such readers to press on with the account in an experimental spirit. Even if we can't get a complete picture of deliberation while excising causal and normative notions, let's see how far we can get. I'll argue that an epistemic model of deliberation can capture enough of deliberation to give us a non-trivial deliberative account of causation. I don't intend the epistemic model to rule out other more detailed investigations of deliberation.

For reasons just given, I have avoided characterising deliberation in normative terms. I have not characterised practical deliberation as working out what *ought* to be the case, or what one has *reason* to do, but as working out what *is* to be the case. This is not because I think practical deliberation can't involve normative reasoning (as I'll explain in section 2.3) but because I intend my account of deliberation to be compatible with a variety of view about practical normativity. Furthermore, normative reasoning doesn't capture what I take to be the most important feature of practical deliberation—that it issues in decisions about how the world *is* or *will be*. Regardless of whether deliberation is meant to issue in a good, justified or rational conclusion, it is meant to issue in *a* conclusion about the state of the world. And this conclusion cannot be merely about normative features of the world. One can work out what ought to be even if one has no intention or expectation of things being settled that way. I am interested in a kind of practical deliberation that aims to settle how the world will be. While forming a normative belief about what ought to be may lead to a practical decision of this form, they are not the same thing.[18]

For similar reasons, I focus on *decision* rather than *intention*—the more common topic in philosophy of action.[19] Intention is often more closely linked to a *commitment* to *make* things a certain way, and so more bound up with norms and causation. I wish to abstract from these as much as possible. Decision, as I'll argue below, is more closely linked with settling things being a certain way. While I take the epistemic conditions I argue for to still apply to intentions, in an attenuated form, decisions are the more straightforward place to start. I will,

[18] Owens (2009) considers cases in which one can rationally judge what to do without settling the normative question of what one ought to do. He also suggests that 'some creatures might be capable of thinking about what to do without deploying normative concepts like 'should' and 'ought''(2009, p. 122). For a similar argument, see Holton (2006).

[19] Moran (2001) takes the *event* of deciding and the *state* of intending to be two aspects of the same phenomenon. I disagree. There are event and state correlates for both: there is the *event* of forming an intention, and also the *state* of being decided.

however, refer to work on intention as it is relevant for decision, and follow decision theorists in also talking of ‘choosing’ and ‘choices’.

I will typically characterise the options agents decide on in terms of states of affairs represented by propositions: an agent decides *that* he will be in London next week, or *that* there will be milk in the fridge.[20] This is in preference to characterising the options in terms of actions: deciding *to* go to London or *to* buy milk. Given that we often do talk of an agent deciding *to ø*, where *ø* is an action, let me say something about how these formulations relate. In many cases, states of affairs and action-based formulations can be converted into each other in a straightforward way.[21] Converting from actions to states of affairs, I might decide *that* I go to London next week or *that* I buy milk: more generally, that I *ø*. The states of affairs formulation can pick out actions to be performed by the agent. Using the first-person pronoun emphasises that such states of affairs may need to include personal indexicals.[22] Converting from states of affairs to actions, I might decide *to* act-so-as-to bring it about that **a**, where **a** picks out the state of affairs.[23] Part of the reason these formulations can be converted is that when we choose actions, we choose them under descriptions. Taking an example from Davidson (1980, ch. 1), I might decide to switch on the light, but not decide to alert the burglar,’ even if these are, in some sense, the very same action. The description

[20] I take propositions to be the objects of propositional attitudes like belief and to be bearers of truth-values. I take states of affairs to be situations that either obtain or fail to obtain and imply the truth or falsity of propositions. I leave to one side the substantive issue of how finely individuated states of affairs should be, and how they relate to the propositions that represent them.

[21] For arguments that they are not convertible, see Schroeder (2011). Schroeder is concerned about downstream effects in normative theories and meta-ethics, as well as semantic issues I do not engage with here. But he accepts that the propositions-formulation can say everything the action-based formulation can, and that is enough for my purposes.

[22] If states of affairs and propositions cannot be indexical, states of affairs and propositions will need to be replaced by theoretical equivalents, of whatever kind are needed to deal with indexical beliefs and indexical propositional attitudes more generally (Perry 1979).

[23] Note that ‘bringing it about’ does not imply that there is an action that is a part of **a** that the agent performs: I might decide that my walls are to be painted blue, without deciding *to* paint the walls (since I do not envisage painting them myself).

we choose actions under can then be invoked in giving a first-person tensed proposition representing the option.

Despite their convertibility, there is good reason to prefer the state of affairs (or some other broadly propositional) formulation, at least for the following project.[24] Firstly, if we are considering the relations between an agent's beliefs and the options she takes to be available, it will be useful if the beliefs and options can be directly compared. This will be easier if the options are characterised in terms of the objects of belief, that is, propositionally. Secondly, characterising options without appealing to actions helps ensure we do not smuggle in features of action that trivialise the explanation of our apparent freedom in deliberation—as we might if we assume that actions are performed freely. Thirdly, as mentioned, since I will use this epistemic model of deliberation to give an account of causation, it will be better to excise causal notions as much as possible from the start. Since the general formulation in terms of deciding *to* uses the notion of 'bringing about', this is a reason against using it. Finally, if we can characterise options in terms of states of affairs that do not mention actions, we leave open the possibility of explaining why certain movements count as actions in a non-circular manner.[25] For example, we might account for why the movement of Georgie's arm counts as an action by appealing to the fact that her decision to move her arm is appropriately related to the movement. While we may be able to give such explanations even if options are characterised in terms of actions, we risk trivializing these explanations. Action theorists, often pursuing other projects, may have reason to prefer the action

[24] One could also hold, for example, that we decide on events, or events under descriptions. The differences between states of affairs and events won't affect the accounts of deliberation and causation I develop.

[25] On a number of views, our capacity for intentional action is closely tied to our capacity to intend or decide. According to the 'Simple View', an agent who intentionally ø's must have 'an intention to ø'—suggested by Bratman (1984) and held by Kaptian (1990; 1994). A weaker view is that intentionally ø-ing requires having an intention, but not necessarily an intention to ø (Bratman 1984). Owens (2009) also argues that our capacity for free action depends on our capacity for decision-making.

formulation: perhaps decisions to act are required for certain kinds of action explanation, or for agents to be self-determining. But provided the reader is willing to accept that we at least decide on actions under descriptions, the conversions given above should be sufficient. For grammatical ease, I will also sometimes talk of deciding *to ø*, shorthand for deciding *that* I *ø*.

I don't assume that deliberation is always conscious. But I will primarily consider intuitions and examples based on conscious deliberation—the kind we are aware of and experience ourselves as involved in—rather than unconscious deliberation—the kind we may, sometimes after that fact, infer ourselves to be engaged in. I choose to focus on conscious deliberation for several reasons. Firstly, it is in conscious deliberation that we typically are aware of our apparent freedom more fully and pay close attention to our mental lives. This makes it easier to identify what epistemic conditions are required for deliberation and apparent freedom. Secondly, cases where we act as a result of conscious deliberation are paradigmatic cases of free agency. Plausibly, it is the possibility of consciously deliberating that ultimately explains why we seem free in cases where we do not act as a result of conscious deliberation. But note that even though the epistemic model is motivated by cases of conscious deliberation, it does not rule out unconscious deliberation. The epistemic model I give can accommodate unconscious deliberation, provided we allow for the possibility of unconscious beliefs as well. There may be difficult cases where beliefs or deliberation are actively hidden or suppressed from consciousness in order to hide tensions between areas of our mental lives. While the conditions I outline below may sometimes hold for these cases, they need not, and these cases should then be treated as aberrant—cases where our mental lives are not sufficiently well ordered for standard conditions of deliberation to apply (see section 2.2c). To highlight the fact that these conditions need not hold of all deliberations, I will sometimes speak of them as conditions on 'reasonable deliberation'.

The conditions I argue for serve two kinds of purposes. Firstly, they aim to identify broad structural features of deliberation that are shared across practical deliberations with different content. In this sense, the conditions aim to *characterise* deliberation, by highlighting its epistemic structure. But I don't intend these conditions to demarcate a mere arbitrary activity. I'll argue that they mark out a coherent and unified phenomenon that plays an important role in our lives. Deliberation serves an epistemic function—it helps epistemically settle what we will do and what will occur, so that we can engage in further planning. For these reasons, I'll argue that these conditions are also *constitutive norms* of deliberation. If we had an activity that systematically flouted these epistemic conditions, it would not rightly be called deliberation.

## 2.2 The Epistemic Model

### 2.2a Epistemic Conditions on Deliberation

In this section, I argue for a set of epistemic conditions that must obtain if an agent is to take an option to be available to decide on in deliberation.[26] By stipulation, an agent must take an option to be available to decide on if she is to deliberate concerning that option.[27] The conditions are necessary for deliberation, but are not each non-redundant. And while I

[26] In developing these conditions, I have relied substantially on the work of Isaac Levi (1980; 1986; 1997; 2000). I discuss the major difference between our approaches in the next chapter (section 3.1c), and other more minor differences as they arise. Tomis Kapitan (1984; 1986; 1989; 1990) uses conditions that are closer to my own, although he takes options to be essentially actions (1986, p. 248). I discuss other differences as they arise.

[27] Note that taking an option to be available does not imply that the option will result if she chooses it, nor the reverse. For this reason, I have avoided speaking of agents 'recognising' an option as available, which suggests the option really is available. For other ways of marking these distinctions, see Smith (2010) and Bok (1998, p. 106).

will argue that the conditions can do interesting explanatory work, they don't aim to give a complete characterisation of deliberation.

An agent takes an option **a** to be available only if:

(a) her deciding on **a** is a serious possibility for her;

(b) she is practically certain that if she decides on **a**, then **a**;

(c) **a** is a serious possibility for her;

(d) it is a serious possibility for her that if she does not decide on **a**, **a** does not obtain.

An agent deliberates only if:

(e) she takes several options to be available;

(f) she takes her apparently-available options to be incompatible.

To begin, let me clarify these requirements. Firstly, these requirements all concern an agent's beliefs—they concern what she takes to be the case, what are 'serious possibilities' and 'practical certainties' for her. I discuss these latter two notions shortly. By stipulation, 'taking' is a weak doxastic attitude, such that the agent has the disposition to affirm the belief when asked, but the belief need not be high-level or conceptualised—it can be merely a low-level awareness or presumption.[28] Secondly, these conditions are inherently first-personal and are indexed to the first-person throughout (such that 'she' takes a *de se* reading). They are not claims about what other agents believe, or what is more generally the case. Thirdly, the conditionals in conditions *b* and *d* are to be read as indicative conditionals. While material

[28] A similar notion is discussed by Kapitan (1986, p. 235). This usage implies that the agent need not have fully articulated concepts of serious possibility, for example, in order to deliberate—*contra* Coffman and Warfield (2005).

conditionals would suffice for my purposes, indicative conditionals more plausibly capture an agent's beliefs.

Conditions *a*, *c* and *d* concern 'serious possibilities'. By definition, a state of affairs *p* is a serious possibility for an agent, if its obtaining is not ruled out of consideration by her other beliefs—that is, if not-*p* is *not* a logical consequence of her beliefs (or any subset of her beliefs, if agents have inconsistent beliefs). Serious possibility is a kind of relativised epistemic possibility. The notion is adapted from Levi (1980; 1986), where it helps explain how a possibility can be assigned a degree of belief zero, and yet still be given consideration. For example, an agent may assign equal probability to a dart landing at any of an infinite number of points in a finite interval of the real number line. This means an agent has a degree of belief zero in each possible outcome. But he still countenances these outcomes as possibilities because each is compatible with his other beliefs.

Kapitan also appeals to a similar notion. Kapitan is concerned with actions that are contingent relative to what the agent *takes it* he believes (1989, p. 35). A state of affairs *p* is contingent relative to a set of states of affairs *S* just in case neither *p* nor not-*p* is a consequence of *S*, i.e., 'that the obtaining of the members of *S* is not sufficient for the obtaining of *p* nor for that of not-*p*' (1984, p. 41).[29] Similarly, we might claim, an option **a** is a serious possibility for an agent, if he *takes it* that if *S* represents any subset of his beliefs not-**a** is not a logical consequence of *S*—implying **a** is logically possible, given *S*. But note that under this formulation, unlike mine, an option may be a serious possibility for an agent, even if it is inconsistent with his other beliefs. Kapitan's formulation allows for more everyday

[29] For arguments on why beliefs should be relativised in this way, see Kapitan (1986, pp. 237, 244; 1989, p. 32; 1990, pp. 130–2).

forms of irrationality. But it only does so by appealing to higher order beliefs. My own formulation aims to avoid appealing to higher-order beliefs, and so is closer to Levi's.[30]

Practical certainty is my own notion, and relates to serious possibility as follows: if an agent is practically certain of *p*, then not-*p* is *not* a serious possibility for her; and if *p* is a serious possibility for her, then not-*p* is *not* a practical certainty for her.[31] Note that having a degree of belief 1 does not imply practical certainty. As noted, *p* can be serious possibility for an agent, even if he has a degree of belief 0 in *p*. In the example above, the agent can have a degree of belief 0 that the dart will land on a certain point, and so a degree of belief 1 that it *won't* land on that point. But its landing on that point can still be a serious possibility for him. So he can have a degree of belief 1 that it *won't* land on that point, yet this not be a practical certainty.

Nor does practical certainty imply having a degree of belief 1, or a degree of belief above a certain threshold—although such characterisations will often work in a rough-and-ready way. The term 'practical certainty' has been used because it suggests that our practical certainties will depend partly on practical matters and on what possibilities we do in fact countenance when reasoning. A belief is practically certain when it is settled for all intents and purposes in an agent's subsequent practical and theoretical deliberations, and when its falsity is taken not to be a serious possibility. This belief can be later revisited, but it is something currently held fixed. A belief can be treated in this way, even if an agent does not

[30] The difference between Kapitan's approach and my own won't matter when I move to more demanding requirements on deliberation in chapter 4.

[31] Levi uses a notion of 'full belief' rather than 'practical certainty', where having full belief rules out incompatible proposition as serious possibilities. If an agent has full belief in a proposition, she takes there to be no serious possibility of its being false. But full belief implies a degree of belief 1, unlike practical certainty.

have a degree of belief 1. What will count as a sufficiently high degree of belief to have practical certainty, on any given occasion, will depend on the needs of the agent regarding her subsequent deliberations—it will be heavily context-dependent, and so not a robust threshold value.[32] Harman appeals to a similar notion in the context of intention: 'forming the intention to do *A* settles in one's mind the question whether one is going to do *A*…in any additional theoretical or practical reasoning, one will be able to take it for granted that one will be doing *A*' (1976, p. 438). Dennett also appeals to a similar combination of epistemic and practical constraints in characterising what options an agent takes to be available (1984, p. 113).

With these points in mind, it is time to consider the requirements more closely. Condition *a* is that the agent's deciding on **a** is a serious possibility for her. Aside from whether **a** obtains or not, an agent takes her actual deciding on **a** to be something that may obtain. In Levi's terms, her beliefs do not rule out her choosing **a** true (1986, pp. 48–52; 1997, p. 77). It is uncontroversial that there must be some sense in which an agent is 'able' to decide on **a**, for **a** to be an available option. If Raamy knows he can't *decide* to swim into a river swimming with crocodiles, because he is so fearful of them that he can't even form the decision, then swimming into the river is not an available option for him—and he won't appear free to decide to. But it's controversial to take an epistemic sense of possibility to capture the relevant ability. One might think that epistemic possibility is neither necessary nor sufficient (with the other conditions) to capture an agent's ability to choose **a**.

[32] For examples of such context-relativity, see Owens (2009, p. 131). If practical certainty is heavily context-dependent, we may need to specify that the belief is held fixed in a sufficient range of potential deliberations.

I will say more in defence of condition *a*'s *necessity* in the next section. But why think an epistemic sense of possibility would be *sufficient*? Ultimately the epistemic model can only be judged as a whole. The question is whether the conditions *together* give a sufficient characterisation of deliberation. In the remainder of this chapter, I'll argue that the epistemic model provides a sufficiently rich characterisation of deliberation to explain our apparent freedom in deliberation (section 2.4). But for now, my motivation for this epistemic condition is that it's our best hope for characterising deliberation in a non-causal fashion, without appealing to primitive agent abilities. Competing conditions are usually either causal, where the agent believes she can *cause* the decision (Joyce 2002, p. 88), or appeal to agent abilities—for example, that the agent believes 'her choosing *is* a conscious effort of *her own*' (Kapitan 1986, p. 249, his emphasis).[33] But if we're to explain an agent's apparent freedom, we do better not to appeal to agential primitives. And if we're to explain causation, we do better not to appeal to causal notions. In the next chapter, I'll also argue against competing attempts to explain our apparent freedom in epistemic terms.

Condition *b* requires that the agent be practically certain that if she decides on **a**, then **a**. For an agent to recognise an option as available in deliberation, she must be practically certain that if she decide on **a**, then **a**. Choosing **a** settles the truth of **a** for her, such that it is no longer something she needs to deliberate or question herself concerning. For example, Shyane should take deciding on there being milk in the fridge to settle the question of whether there will be milk in the fridge. The state of affairs is now a fixed point in what she

[33] One could also appeal to desires—an agent feels free to make a decision, just in case her decision is in keeping with her desires. But if we're to explain why an agent feels free to decide in different ways, we need to appeal to more—perhaps that her desires are conflicting, or may change, or she doesn't know her current desires—so that her desires leave more than one option 'open'. I suggest we'll have to crucially appeal to epistemic notions as well. I consider some of the ways desires (or other evaluations) can play a role in deliberation under the epistemic model in section 2.3, and consider other desire-based accounts in section 3.4.

believes to be the case, such that she can make further plans and preparations without having to revisit the question.[34] Shyane may still renege on her decision and revisit the deliberation. But she cannot take reneging to be a serious possibility as she decides.[35] Note that this does not imply that a decision *is* a belief, or has belief as a component part: decision merely *implies* belief in the outcome.

Conditions *a* and *b* imply condition *c*: that **a**, the state of affairs obtaining, is a serious possibility for the agent—something that may come about for all she believes. If deciding on **a** is a serious possibility, and deciding on **a** epistemically settles the obtaining of **a**, then **a** must also be a serious possibility. So if Shyane is to deliberate concerning two available options, such whether to have milk in the fridge or not, then both states of affairs must be serious possibilities for her: her beliefs about the world can't fix for her whether there is milk in the fridge or not. Because this condition plays a particularly important role in the account, I have included it as a separate condition. Condition *c*, like *a*, is controversial; defending it will take up much of the next section.

Condition *b* is also controversial: that deciding on an option epistemically settles its obtaining. I'll say something in defence of it here. The condition has it adherents.[36] It also helps explain what options are taken to be available in deliberation. Consider van Inwagen's case of a man in a room with two doors, believing one to be locked, the other unlocked, but who does not have beliefs that settle which is which (1983, p. 154). Van Inwagen claims that

[34] For more on the epistemic function of intention in allowing for further deliberation, planning and coordination, see Bratman (1984).

[35] For discussion of how new evidence may bring an agent to re-enter deliberation, without being part of deliberation, see Owens (2008, p. 263).

[36] Grice (1971), Harman (1976), Levi (1986, 1997) Kapitan (1986, 1989) all endorse such a condition. Slightly weaker conditions are defended by Hampshire and Hart (1958), Ginet (1962), Taylor (1964), Schick (1979) Holton (2006) and Setiya (2008).

the man cannot sensibly deliberate about which door to open—and uses this to support his metaphysical account of free choice. Condition *b* secures the same result without requiring metaphysical possibilities. Even if the man takes door 1 being unlocked to be a serious epistemic possibility, and door 2 being unlocked to be a serious possibility, he is not practically certain that if he decides on door 1 being open, door 1 will be open, and likewise for door 2. So he does not take these options to be available.

One might worry that condition *b* is too strong: an agent need not be certain **a** will obtain in order to recognise **a** as available. He need only consider it *likely* that **a** will obtain, or take his deciding to raise the probability of **a**. For example, it might seem golfer-Dan may recognise making a difficult putt as an available option, even if he is uncertain of whether he will succeed. In response, consider what happens when we try to apply the language of 'decision' to such cases. While we might speak of the Dan as *intending* to make the putt, as intending to *try* for the putt, or as *deciding* to *try* for the putt, we would not naturally speak of him *deciding* to make the putt. Similarly, we would not usually describe a player as *deciding* to win a lottery, or *deciding* to win a coin toss, even if that is what they intend, try for, or think they raise the probability of. We reserve the term 'deciding' for an outcome that is settled, an outcome that we hold fixed in our own further planning and deliberation, and license others to hold fixed as well.[37] And the reason for this is not simply that we happen to use words in this way—these words pick up on important distinctions, distinction that themselves map to features of our practices. It makes an interesting difference whether a choice settles for one that the

[37] Separating intention and decision in this way also allows us to accommodate competing intuitions concerning whether an agent can intend an event that he cannot reliably bring about. Under my preferred view, the agent can intend the state of affairs (because of the commitment involved) but not decide on it. The etymology of 'intend' and 'decide' bears out this point (although I don't place much weight on this). 'Intend' derives from the Latin *intendere*, whose meanings relate to stretching and extending, such as when one bends oneself towards a certain end. 'Decide' comes from the Latin 'decidere' and has meanings related to cutting off, settling or determining something in question or doubt (*Oxford English Dictionary*, Murray and Burchfield (eds.) 1933).

option will occur, or whether it merely settles that the option will be sought. If an option is settled, we can justifiably inform others of its occurrence, and plan around the occurrence ourselves.[38] If it is merely sought, we cannot. Taking into account possible failures of performance, in particular, is essential to good planning. Moreover, note that an agent can still *decide* that the probability of an outcome be raised, or to try for an outcome, even if an outcome is too unlikely to be decided on directly. We can still preserve the same conditions on decision, with no loss of generality. If we do wish to extend this account of practical deliberation to cover options that are too unlikely to be decided on, what we should do is alter the option-descriptions, rather than condition *d.* The option-descriptions could be deciding on a state that raises the raise the probability of **a**.[39]

One might still wish to argue that there is a certain threshold degree of belief at which the outcome counts as decided upon. Much of what I say in the following is compatible with such a view. But it is difficult to imagine that there is any robust exact value at which this occurs—it will be heavily context-dependent. Because the threshold value must be sensitive to the needs of the agent's subsequent deliberations, we will be lead to something like the pragmatic account of certainty that I prefer: a belief is practically certain when it is settled for all intents and purposes in an agent's subsequent practical and theoretical deliberations.

[38] Grice make a similar point: in stating 'I will do A', an agent 'is ordinarily held both to have given others a justification for planning on the assumption that he will do A and to have committed himself, on pain of being convicted of insincerity, to planning on that assumption himself' (1971, p. 4). See also Harman (1976, p. 434).
[39] If the epistemic conditions are to explain why options that have been rejected are no longer available, we can't weaken condition *b*, or rejected options will still count as available. A second reason to alter the option descriptions rather than the conditions is that the incompatibility required by condition *f*, considered below, concerns decided states, not the intended ones. Unless there is some incompatibility in *trying* for (or raising the probability of) two incompatible outcomes, there is no associated irrationality in intending both of them. While there are situations in which trying for (or raising the probability of) two incompatible options may undermine one's ability to achieve either, if this is not the case, it not irrational to intend both.

Condition *d* requires **a**'s not obtaining to be a serious possibility for the agent, given she does not decide on **a**. In other words, the agent can't be certain that **a** will obtain, regardless of how she decides. For example, an agent cannot decide that 2+2=4, or that the sun will rise tomorrow, if she takes these to be settled, regardless of what she decides. This would prevent her deliberation from having any epistemic bearing on what option obtains, and agents must take their decisions to be epistemically relevant for how the world goes. I take condition *d* to be uncontroversial as a necessary condition on recognising an option as available, and I will say nothing further in defence of its necessity.[40] Causal decision theorists may also require the agent to take her decision to be causally efficacious. But this is no reason to deny this minimal epistemic condition.[41]

In addition to these four conditions on an agent taking an option to be available, there are two conditions on an agent deliberating:

(e) she takes several options to be available;

(f) she takes at her apparently-available options to be incompatible.

If an agent is to deliberate, she must take more than one option to be available. And she must take her apparently-available options to be incompatible, in the sense that she takes any one option to rule out the possibility of any other. Options might be taken to be

[40] One could also allow the option to be taken as available, but maintain that deliberation could not take place, since there would be no alternative options available (Levi 1986, p. 60). I take this to be a notational variant and not a substantive disagreement. Note that condition *d* is redundant if one assumes that an option can only be available if another incompatible option is available. I've allowed that an option may be available in this situation, but that an agent may not reasonably deliberate (by conditions *e* and *f*).

[41] Kapitan introduces a stronger condition—that the agent takes it he will act *if and only if* he chooses to (1986, p.234). This is plausible, but only if options are actions. If options are states of affairs, the condition is too strong as a condition on intention (as it is for Kapitan). I can intend a state, even if I'm aware that the state may be brought about by other means. But ultimately my conditions on decision will have similar consequences as Kapitan's, since I also require the agent to take some of her available options to be incompatible.

incompatible because of nomological, causal or logical grounds: for example, an agent might deliberate between being in Manhattan or Brooklyn tomorrow at 8pm, presuming she cannot be in both places at once. Or she might deliberate between being on the London train or not being on the London train, taking these options to logically exclude one another. Together conditions *e* and *f* imply that an agent takes it that several incompatible options are available.

Why think that alternatives must be taken to be incompatible? Say Alice deliberates about whether to hit the top two-thirds of a dartboard, or the bottom two-thirds. Her options are compatible—the dart may land in the middle third, which is compatible with both options. Let's further specify the case and say that Alice can't decide to hit any single third—she's not a good shot, so these (incompatible) options aren't available to her. But there are other incompatible options available to her. Alice can decide that she aim in an upwards direction, or decide that she aim in a downwards direction. So we haven't ruled out Alice deliberating by requiring options to be incompatible. Are there reasons for Alice to prefer to demarcate her options in this way? In some cases, it may help Alice focus on the differences between her options, rather than the similarities—if aiming in an upwards direction rules out hitting Alice's disliked bottom third, this may be a reason for her to prefer it. In these cases, while there are reasons to carve available options into incompatible ones, I suspect they are not decisive.

But what if more finely demarcated incompatible options are available? Say Alice is a good shot, and can decide to hit any single third. Or say Alice can choose any of the items on a café menu. In these cases there are stronger reasons for carving up options into incompatible ones. Say Alice is deliberating between a chocolate desert and a tart, where there are several of each on the menu. If there is also a chocolate tart on the menu, this is not a good way for Alice to carve up her options. Why? Because having chosen a tart, she will still have to decide on whether to take a chocolate one or another, and conversely for deciding on a chocolate desert—and precisely the same reasons, concerns and desires will be in play both times. In addition, it may be difficult for Alice to decide between chocolate or a tart. Yet there is no need to do so if she chooses the chocolate tart. Alice does not need to determine whether she chooses the chocolate tart in virtue of its chocolaty or tarty qualities. These are strong reasons in favour of carving up options into incompatible ones. Finally, note that allowing for partially overlapping options will not affect my arguments below—it just makes them more difficult to state. It will still be in terms of the incompatible parts of the option alternatives that we make sense of deliberation and its epistemic function.

What if the options are in no way incompatible? Say Alice deliberates between the mango tart, and the chocolate truffle, where taking one doesn't rule out taking the other. Once again, it seems Alice hasn't demarcated her options well. If she decides to have the mango tart, this in no way settle her *not* having the chocolate truffle, and she will have to continue to deliberate with the same kinds of considerations in play. And if *different* considerations are in play, it's hard to see why we should call it a single deliberation. If Alice's choosing her first desert is in no way relevant for her second choice, it's hard to see why this is not simply two deliberations—one for her first choice, one for her second—and incompatible options are again in play. Deliberation must be between (at least partly) incompatible options.

Altogether, we now have six necessary conditions on taking an option to be available in deliberation. Only four of these are non-redundant: *a*, *b*, *e*, *f*. As already noted, *a* and *b* imply *c*. Furthermore, *d* is implied by the other conditions. Here's why. Deliberation is between apparently incompatible alternatives (*e* and *f*), where the decisions settle the states obtaining (*b*). So if an agent decides on an option taken to be incompatible with **a**, she must take not-**a** to obtain. So an agent must take it to be a serious possibility that if she does not decide on **a**, then not-**a** (condition *d*). It's worth keeping in mind that these are relatively weak requirements on deliberation, and allow an agent to be unsure about her abilities and the state of the world—only one of these requirements requires certainty, and only practical certainty at that. Note also that my claim at this stage is only that these are necessary requirements. They may not be jointly sufficient, and there may be other necessary requirements.

But even though these epistemic conditions are minimal, they already do important explanatory work. Firstly, they capture an agent's unsettledness about her *decision* in deliberation. Conditions *a, e* and *f* imply that she takes several decisions to be possibilities for her, and that her beliefs do not settle which decision she will make. Secondly, they capture an agent's unsettledness about subsequent *states of the world.* Conditions *c*, *e* and *f* imply that the agent's beliefs do not settle which of several incompatible states of affairs will obtain. She is uncertain of the world. Thirdly, they capture how the state of the world depends (epistemically) on the agent's decision. Conditions *b* and *d* imply that an agent's decisions settles what state of affairs obtains, where this is otherwise unsettled. So these conditions capture the way in which an agent's decisions are *epistemically relevant* to how the world goes, the proof of which follows.

*Proof of epistemic relevance:*

Where **a** is taken to be an available option by a deliberator S:

S is practically certain (S decides(**a**) → (**a**)) (*b*)

S decides(**a**) → S is practically certain (**a**) (conditional belief)

S is not practically certain (not(S decides(**a**)) → (**a**)) (*d*, def. of practical certainty)

not(A decides(**a**)) → S is not practically certain (**a**) (conditional belief)

∴ S decides(**a**) is relevant to A's practical certainty in **a**

S deliberates between some options she takes to be incompatible (*e*, *f*)

∴ S's decisions are epistemically relevant to states of affairs she takes to be incompatible.

### 2.2b The Impossibility of Predicting Your Own Decisions

If we accept the above epistemic conditions on deliberation, we can derive the following result: a deliberating agent cannot be certain of what decision she will make.[42] Call this the *impossibility thesis*.

*Proof of the impossibility thesis:*

Where **a** is taken to be an available option by a deliberator S,

S is practically certain S decides(a)

→ S practically certain ((S decides(**a**) → **a**)∧(S decides(**a**)) (*b*)

→ S practically certain (**a**) (indicative conditional)

→ not(A takes **b** to be a serious possibility) (definition)
(for any **b** taken to be incompatible with **a**)

→ not(A takes **b** to be available) (*c*)

∴ S takes no options taken to be incompatible with **a** to be available

∴ S does not deliberate (by *e*,*f*)

If an agent were to become practically certain of her decision **a** while deliberating, this would imply (by condition *b*) that she is practically certain of **a**. Condition *b* requires that choosing

[42] Claims that the deliberator must be ignorant of the decision or the outcome during deliberation are made by Shackle (1958), Hampshire and Hart (1958), Ginet (1962), Taylor (1964), Perry (1965), Prior (1968), Harman (1976) Jeffrey (1977), Schick (1979), Dennett (1984) and Gilboa (1999). Pereboom (1995) also raises this possibility. Jones (1968) and Kapitan (1986, 1989) explicitly analyse our freedom in deliberation in terms of ignorance. For objections, see Cox (1963), Stocker (1968), Velleman (1989a; 1989b) and Joyce (2002; 2007). Note that I am not making the further claim, as Ginet, Taylor and Prior do, that there must be a similar ignorance with respect to other people's decisions—for arguments against this claim, see Perry (1965). Levi argues for a somewhat stronger thesis under the slogan 'deliberation crowds out prediction', requiring that agents cannot have *any* credence concerning their decision during deliberation (1986, ch. 4; 1997, chs. 2, 4; 2000), tracing his arguments to Spohn (1977). I consider Levi's view in the next chapter, section 3.1c.

an option settles the option's obtaining. This implies (given minimal rationality) that for any other option **b** taken to be incompatible with **a**, she does not take **b** to be a serious possibility. So by condition *c*, **b** cannot be taken as an available option. Condition *c* requires options to be serious possibilities. The above holds for any option taken to be incompatible with **a**. So no apparently incompatible options can be taken to be available. Since more than one incompatible option is not taken to be available, there can be no deliberation (by conditions *e* and *f*). While prior to deliberation the agent may predict what she will decide, she cannot maintain this prediction if deliberation is to take place.[43]

Before providing further defence of this thesis, let me say something about the kind of impossibility involved. Recall, these requirements on deliberation were put forward partly as normative requirements on deliberation. These conditions may sometimes be broken, for agents who are less than fully rational. But the failure to live up to these normative standards of deliberation, if widespread, would prevent the activity being properly characterised as deliberation. For example, we might accept the conceptual possibility of a neurotic who seems to "decide" everything twice, in the sense that he engages in two successive deliberation-like activities for every choice (without abandoning or forgetting his first "decision"). But if his second "deliberation" contravenes the epistemic conditions in a sufficiently wide range of cases, the activity should no longer properly count as deliberation.

The impossibility here is similar to the impossibility involved in holding plainly contradictory beliefs. It is in some sense psychologically possible for an agent to hold plainly contradictory

[43] If one is dissatisfied with the pragmatic approach to characterising certainty and serious possibility above, one could replace these notions with others and still derive the same result, provided conditions *a* and *d* were related in such a way that believing an option will be chosen makes other incompatible options no longer serious possibilities.

beliefs, beliefs whose incompatibility is clearly apparent when the beliefs are brought to the agent's attention. But holding plainly contradictory beliefs not only goes against norms of rationality, but, if wide-spread across an agent's belief-forming practices, prevents her states being well-characterised as beliefs. Precisely because plainly contradictory beliefs flout epistemic norms so flagrantly, when it comes to interpreting what a particular mental state might be, we are left without our usual guidance. If we think that mental states like beliefs and decisions are accessible through this kind of interpretation, we are led to impossibility theses regarding beliefs and practical deliberation.

One might think the impossibility thesis is too weak to be of interest. Could we ever be completely convinced of what we'll do before we do it? I'll consider some cases in the next section. But, for now, recall the kind of certainty involved is only 'practical certainty'—holding the state fixed and settled as far as one's subsequent deliberations go. It is not complete certainty. We do seem to encounter this kind of certainty in our ordinary lives—when we take it for granted that the store will be open at the usual time, that we'll act as we've decided, that the sun will rise.

Why should we believe the impossibility thesis? In the next chapter I'll consider competing models of deliberation from James M. Joyce and David E. Velleman that directly challenge the incompatibility thesis. Here I'll consider other general arguments in its favour and against it. Note that the arguments I give to defend it also act as defences of requirements *a* and *c*—that apparently-available options and decisions are serious epistemic possibilities. Both of these conditions are typically rejected when attacking the impossibility thesis.

To begin, it does seem to be a deep component of our understanding of deliberation, that if an agent knows how she will decide, and what will result from her deliberation, deliberation is out of place. Say Matt becomes convinced that he will decide to call his mother this Sunday, and that he will call her this Sunday. Is this something he can then sensibly deliberate concerning? His calling her is something he has already settled, and can hold fixed in his subsequent practical and theoretical deliberations. He can already send her a quick message letting her know he'll call. He can tell other people he won't be free at that time. He can engage in more extensive planning, and think about what they might talk about. As far as Matt's calling her goes, there seems nothing left for him to decide. Imagine Mat reporting being certain of the outcome of a deliberation, yet still deliberating: 'I'm going to call mum on Sunday night…but I'm still not sure whether to call her'. This sounds decidedly odd. Our difficulty in conceiving of such scenarios reflects a deep feature of deliberation.

Here's another reason for thinking deliberation requires uncertainty. Making a decision is often taken to imply *forming* a belief about what one will do (Hampshire and Hart 1958; Pears 1968; Grice 1971). If making a decision implies forming a belief, this explain why there is typically an equivalence between reporting what one has decided to do, and what one is going to do—both can be expressed by a phrase like 'I'm going to call my mum on Sunday night'. One's decision and belief get settled together. If so, deliberation would require one's belief *not* to already be settled.[44] There are also accounts that take decisions to *be* beliefs about what one will decide or do (Velleman 1989a; 1989b; Joyce 2002; 2007; Ismael 2012).

[44] A number of authors also note an analogy between deciding and realising, discovering and being surprised (Ginet 1970, p. 177; Schick 1979; Sorensen 1984). One cannot be certain of A immediately prior (or while being) realising, discovering or being surprised by A.

As I'll argue in the next chapter (section 3.2), these views also plausibly require that an agent doesn't already have the relevant belief if she is to deliberate.

We can get a deeper understanding and defence of the impossibility thesis by considering the epistemic and doxastic functions of deliberation. Its doxastic function is to allow us to become certain of what we will do. We can then rely on this certainty in our subsequent deliberations and planning, and license others to rely on it as well. Belief in the decision allows the agent to infer the outcome, in a way she was not able to otherwise. As Kapitan puts it, '[i]f deliberation aims at settling the mind upon a course of action, then a deliberator must be partly unsettled prior to decision' (1990, p. 127). Harman makes a similar functional argument (1976, p. 438): [45]

> …forming the intention to do A settles in one's mind the question whether one is going to do A. It can be useful to settle that question because, having done so, one will no longer need to consider whether or not to do A and one can turn one's mind to other issues…in any additional theoretical or practical reasoning, one will be able to take it for granted that one will be doing A.

While deciding may not give one complete certainty that the event will occur, it provides as much certainty as predicting the decision.

There is a closely related *epistemic* function of deliberation: coming to justified certainty. Carl Ginet is particularly explicit concerning this function: 'the whole point of making up one's mind is to pass from uncertainty to a kind of knowledge about what one will do or try to do' (1962, p. 52). The epistemic requirements on deliberation I've argued for can explain how we can come to justified belief, not just certainty, of what we will do. Agents often have direct

[45] Similar arguments are made in Ginet (1962), O'Shaughnessy (1980, p. 297), Holton (2006) and Levi (1986, p. 86).

knowledge of their own decisions. So by forming decisions, agents can come to know what options will obtain. Whether an agent has knowledge or mere certainty will depend partly on whether she justified in believing that the decision settles the option.[46] In general, if an agent is so justified, she can come to know of an option's obtaining by making a decision. And if she *comes* to know her the state that will obtain through deliberation, this requires her not knowing the state beforehand. There may of course be some cases where an agent merely believes (but does not know) what option will obtain. In such cases, satisfying this epistemic function of deliberation does not require failing to have beliefs beforehand, but only failing to have knowledge. But unless we habitually form unjustified beliefs in what option will result, satisfying the epistemic requirement will usually require the agent to fail to have beliefs beforehand. So deliberation having an epistemic function also provides further support for the impossibility thesis.

A competing conception of the function of deliberation is that it fixes the will of the agent upon a certain course, helping to overcome phenomena such as weakness of will. Fulfilling this function might seem compatible with the agent being certain of how he will decide, and so would constitute an argument against the impossibility thesis. Stocker, for example, argues that 'the point in deciding to do something that one knows one will do may not lie in the further traffic with reason or knowing, but rather with the will, desire, or conation' (1968, p. 68, emphasis removed). However, taking there to be an important epistemic function to deliberation does not prevent us acknowledging conative functions as well. The epistemic model requires the agent to believe that the option will occur, given she decides on

[46] Further requirements may be needed as well. But condition *b* will likely still play an important role in accounting for such knowledge.

it. Fulfilling this epistemic condition may well require various conative resources to be in place, such as an agent being motivated to act as she decides.

By appealing to the epistemic function of deliberation, we go a little deeper than merely claiming it to be part of the nature of deliberation that several options are recognised as serious possibilities. But we are not committed to the unintuitive claim that the only benefits of deliberation are of this epistemic type. Deliberation may also be a way of working out a rational or reasonable choice. Nor are we committed to the claim that other kinds of reasoning are unimportant in this context. Once a decision is made, one can become clearer about one's reasons, and so better placed to justify or explain one's decision to others. One can consider the benefits involved or one's justification and perhaps regret or even renege on the original decision—and so re-enter deliberation. Recognising an epistemic function to deliberation does not prevent us recognising other functions to deliberation and other forms of practical reasoning.

### 2.2c Decision Predictions by Others

So that we can further examine the impossibility thesis and its consequences, let's consider some cases where decision predictions actually occurs—cases where someone else predicts an agent's decisions. Even if the agent herself cannot predict her decision while deliberating, nothing I have said prevents another person doing so (including her earlier and later selves).[47] If the agent trusts the predictor's skills, she may have good evidence of what she will decide. This is the kind of situation in which the epistemic limits on deliberation are manifest.

[47] Some unusual circumstances may be required for an agent to predict his decision, and then forget that prediction at the time of deliberation. For examples, see Goldman (1970, ch. 6).

Consider a case. Jacinta is deliberating about whether to spend her week off at home or travelling on the coast. If a friend of hers tells her that she will decide to stay home, there are a number of responses Jacinta could make that would be consistent with the impossibility thesis. Firstly, Jacinta might take the apparent 'prediction' as a recommendation—her friend thinks she should stay home (or wants her to) without its being, on the basis of evidence, what she will do. Perhaps her friend is trying to influence her. Secondly, Jacinta may for this or some other reason mistrust her friend's prediction, disregard it or ignore it, and thereby not form a belief based on it. She might consider it poor evidence of what she will decide. Thirdly, Jacinta might forget or more actively repress the prediction, not through explicit mistrust her friend's ability or sincerity, but through carelessness or self-deception. She might not want to face up to being the type of person that stays home. Or, fourthly, Jacinta might accept that the prediction *was* an accurate prediction before it was expressed—but now that she has heard it, the decision context has changed, and she now has new information that invalidates the previous prediction. Jacinta might take it she has evidence about what decision she *would have* formed, but not evidence of what decision she *will* form. A good predictor should take such changes of context into account, and make their prediction for the decision context the deliberator will be in once they've heard the prediction.[48] But Jacinta may not believe her friend to be as skilled predictor as that, able to take into account the way expressing the prediction changes the context.

[48] Hilary Bok claims the only way one could accurately predict an action is if the action is one that would be performed *regardless* of the prediction (1998, p. 82). This is false. As Goldman notes, a predictor can consistently take into account the effect of his prediction in the prediction itself and, under certain continuity conditions, there will always be at least one stable prediction available (1970, p. 183–5). Perhaps Bok's concern is that if there is more than one prediction that would be accurate, were it to be expressed, they cannot be predictions. But whatever they are, they function as predictions once made.

But what if there were a predictor that skilled? Someone with infallible access to the future, or able to calculate the effect their prediction will have on an agent's action and determine a fixed point where the expressed prediction matches the agent's subsequent action. Let's call such a predictor 'Sandy', an anthropomorphised 'Book of Life' of the type discussed by Alvin Goldman (1970, ch. 6). One day ordinary Mark meets Sandy. Mark has lots of experience with Sandy, watching her make predictions of himself and others, testing the results, and observing that all her predictions come true. After being sceptical for some time, Mark begins to trust Sandy, taking whatever prediction she makes to be good evidence about what he will decide. What then?

One line of thought is that even if Mark trusts Sandy and her prediction, there is still something left for him to work out: what to decide. Even if he believes the future is epistemically closed, he can still take himself to be a free participant in its creation—he knows his decision is part of what makes the future the way it will be. And so Mark's apparent freedom in deliberation remains, even if he is certain of what he will decide. So the impossibility thesis is false.

I want to resist such a line. We can keep the impossibility thesis. If we do so, what in fact happens will depend on what we hold fixed: which features are part of the setup of the case, the independent variables, and which are dependent variables. Say we hold the infallibility and sincerity of Sandy fixed. In this case, if Sandy predicts that Mark will decide to stay home, rather than travel, and informs him, then Mark will indeed decide to say home. This much is guaranteed. But Mark may still satisfy the epistemic conditions on deliberation. He may mistrust her prediction after all, and perhaps foolishly take it that she has not taken into account the new context they are in. So he won't form the relevant belief. Or he might be irrational or forgetful, and so not form the relevant belief, even while trusting Sandy in general. In any of these cases, the impossibility thesis holds.

Alternatively, we might hold both Mark's rationality and his trust in Sandy fixed—such that he appropriately forms beliefs based on Sandy's testimony. If Sandy tells him he will *decide* to stay home, he will believe he will. If the impossibility thesis is true, Mark can't then reasonably deliberate on staying home. He might still end up staying home—but this won't be through deliberating and deciding. So Sandy turns out to be unreliable after all. If the impossibility thesis is true, Mark is rational, and the prediction is expressed, understood and believed, then a prediction that Mark will decide on **a** turns out to be undermining.[49]

However, if Sandy's prediction is merely about what Mark will *do*, not about what he will *decide*, another option is available. Sandy can keep her infallibility. And Mark can still keep his trust in her and his rationality. Mark does stay home and believes he will. But Mark's ability to reasonably deliberate and decide will still have to go. Mark will also, I will later argue, lose his apparent freedom in deliberation. There are different attitudes Mark may take to his staying home. He may still consider it his action, something he *does*, but an instinctual one over which he has not deliberated. Or he may consider it to be not his action at all, and take himself to become merely an instrument for the world.[50]

[49] Sandy might also make a conditional prediction: that if Mark decides at all, he decides on **a**. Conditional predictions are not similarly undermined.

[50] Robert Silverberg's *The Stochastic Man* (1975) and other works of science fiction attempt to describe what such an attitude may be like. Related attitudes are also explored in the context of time travel (Smith 2005).

Sandy may even be malicious, and (incorrectly) predict Mark's decision or (correctly) predict Mark's action in order to prevent his deciding. She may be aware that her prediction is part of the causal history for why he acts as he does, and why he doesn't deliberate. But even if such manipulation occurs, it doesn't imply that Mark incorrectly takes Sandy's prediction as a *reason* to decide. Unless alternative options are taken to be available, there can be no practical deliberation in which reasons might play a role. Nor does the case imply that Mark makes the fatalistic error of assuming that his decisions and actions come about *regardless* of what he does. None of the consistent alternatives I have described require the agent to think his decision-making is irrelevant to what occurs. While it is unfortunate that agents can be manipulated by being told the truth, this is a standing possibility for life, and we should not be required to adopt conditions on rationality that make us immune from it.

When beings like Sandy predict our decisions, there are solutions available that are perfectly consistent with the impossibility thesis. It can be hard to see them. And some of these solutions may look *prima facie* unlikely. But these are not sufficient reasons to reject the impossibility thesis. While supernatural cases are good for laying out the logical implications of a view, we need to be careful when appealing directly to our intuitions regarding them. We live in a world without oracular entities like Sandy. To include such entities, we have to change the epistemic and causal structure of the world we conceive of drastically. It is no surprise that we then have difficulty keeping track of what else might change in such a world, particularly regarding our attitudes and behaviour. In dealing with non-standard cases, rather than appeal to one-off intuitions about what will occur, we do better by laying out consistent solutions and principles and seeing how they accord with our theories elsewhere. We might not be able to keep all our *prima facie* intuitions. But given we have limited experience with beings like Sandy, and given the radical changes involved from cases of ordinary deliberation we are familiar with, this is not a surprising result.

### 2.3 Normative Reasons in Deliberation

At the start of this chapter, I characterised practical deliberation without appealing to normative reasons about what one should or ought to do. I will now say something about how normative reasons for action can feature in deliberation, and even be essential, while retaining the epistemic model and the impossibility thesis. When I qualify reasons as 'normative', I intend to distinguish them from both explanatory reasons as well as from more neutral 'deliberative considerations', where these are just features considered in the course of deliberation. By stipulation, a consideration being a normative reason implies something stronger than an agent merely treating it as a consideration in favour of or against a proposal. Normative reasons carry normative force, requiring, for example, that someone in relevantly similar circumstances *should* take the same consideration to count in favour of the action. Considerations, even considerations I act in light of, need not have or be taken to have normative force.[51]

To begin, lets consider a normative-based objection to the impossibility thesis. I argued above that if an agent is practically certain of **a**, or that she will decide on **a**, she can no longer deliberate concerning **a**. But one might argue that there is still something important left for her to decide: what to do or whether **a** *ought* to obtain.[52] Working out what one ought to do might be thought an essential part of practical deliberation. If so, since prediction

[51] One might take all considerations appealed to in deliberation to be or be treated as normative reasons. I have no objection to such a view. I simply note the bare logical possibility of more neutral kinds of considerations.
[52] For this section, I will switch to an action-based characterisation of options, in keeping with how action theorists characterise the normative reasons involved.

leaves the normative question undecided, an agent can practically deliberate while predicting her decision. Christine Korsgaard defends such a view, claiming that even if we predict our own behaviour, we must still 'decide the case on its merits' (1996, p. 95). Richard Moran suggests a similar line, by distinguishing between practical and theoretical uncertainty (2001, p. 56). It seems one may be theoretically certain of what one will do, yet practically uncertain—and so still deliberate.[53]

In order to respond to this challenge, we first need to be clear on how normative deliberation might feature in practical deliberation. Firstly, we might take practical deliberation to be working out what one ought to do, regardless of what one believes about one will do. In this case, it is clear that predicting what one will do does not prevent practical deliberation. But then it seems we are simply changing the topic. If working out what one ought to do carries no implications for what one believes one is will do (or vice versa), this type of deliberation is merely a type of theoretical deliberation about normative reasons, and, without further argument connecting it to action, we have no reason to characterise it as *practical* deliberation at all. The impossibility thesis could still apply to practical deliberation. Similarly for the suggestion that settling the normative question and settling what one will do are entirely independent and separable parts of practical deliberation. Not only is this an implausible view, but the impossibility thesis would still apply to the non-normative part of deliberation.

Secondly, one might suggest that the active *considering* of normative reasons is essential for practical deliberation, and this is why practical deliberation can take place even while

[53] Joyce (2002) also suggests decisions must be formed in response to normative judgments (see section 3.4), but without using this as an argument against ignorance conditions.

decisions are predicted. Considering reasons might be essential because doing so is relevant to the normative part of deliberation (working out what one ought to do), or because doing so is independently important (perhaps it allows us to justify our decisions and reason with others). However, on either approach, this requirement is no threat to the impossibility thesis. If the benefit of such active considering holds independently of the normative part of deliberation, then one still has reason to require active considering, even if freedom in deliberation is to be explained in epistemic terms. In itself, this approach puts no pressure on the impossibility thesis. If such active considering is only helpful for reaching a normative conclusion to deliberation, we are back with one of the above views—such active considering is not essential to the part of deliberation that settles beliefs in action.

Thirdly, one might hold the rather unusual view that practical deliberation is about settling what one *ought* to do, but this only goes via settling what one *will* do: in virtue of deciding what I will do, I thereby decide what I ought do. While this does establish a connection between the two components of deliberation, it suggests that practical deliberation consists in post-hoc rationalisations. It is surely not what theorists like Korsgaard have in mind.

Finally, we might claim that we settle what we will do via considering what we ought to do. This is the most plausible way in which normative reasoning might be a necessary component of practical deliberation. This is surely the view that Korsgaard has in mind. However, this view is perfectly compatible with the claim that when it comes to deliberating on how to act, we are limited in epistemic ways—to what options are serious possibilities for us. Normative considerations are then only relevant to practical deliberation insofar as they concern epistemically possibilities. If I am certain that I will not *ø*, then normative reasons concerning *ø* are not relevant to my decision. Reasoning normatively in practical deliberation will then imply that different options are epistemically possible. But the normative part of the deliberation will not independently determine which options are available. Instead, the epistemic model will explain the kind of freedom we experience in normative deliberation, of the type that is relevant to practical deliberation.

The lesson of all this is that even if we take normative reasoning to imply a type of apparent freedom or openness, this doesn't imply that normative reasoning is the *source* of this apparent freedom or openness. Korsgaard intends to explain our sense of freedom in terms of normative reasoning. She claims that in deliberating we act 'under the idea of freedom' (1996, p. 97) and that it is our reason-responsive reflective nature that 'gives us a choice about what to do' (ibid., p. 96). But taking normative reasoning to be an essential part of deliberation is compatible with the freedom we take ourselves to have in practical deliberation having its source in epistemic features. Attributing a kind of freedom to our normative reason driven deliberations does not show the impossibility thesis to be false.

However, a problem arises with this picture of how normative reasoning is part of practical deliberation. I'll spend the remainder of this section addressing it. Say Grant is certain, in general, that he'll decide and act as he ought. And say he becomes certain of what he ought to do. He really ought to take his daughter to the park. He thereby becomes certain that this is what he is going to decide to do and do. His normative reasoning allows him to predict his decision. But then it seems his certainty about what he *ought* to do gets in the way of his actually deliberating and deciding what *to do*. According to the epistemic conditions I've argued for, Grant no longer takes several options to be available. So he can no longer reasonably deliberate and decide between them. More generally, even if normative reasoning doesn't allow Grant to pick out a single option as what he ought do, the options left available to him are options that normative reasoning can no longer help him decide between.[54] If normative principles should play a non-trivial role in the agent making his final choice, epistemic conditions on deliberation seem to land us in trouble.[55]

Levi's response to this problem is to claim that deliberators must not, at the moment of decision, believe they will choose rationally—they must give up what he calls the 'smugness assumption' (1997, p. 31).[56] He argues that this assumption is never needed, at any point in deliberation, for agents to apply principles of rationality to their choice. And he argues that it is plausible than an agent will not make this assumption, given he might fall victim to weakness of will and decide out of step with what he believes he ought to do (1997, p. 33). Gilboa's response is similar. He argues that 'regarding others, one may "believe" or even "know" that they make choices in a consistent way with past ones, so that their utility functions may be used for prediction. As for one's own self, however, no similar consistency axiom can be known or believed' (1999, pp. 169–70).

But Levi's response is problematic. Firstly, it requires systematic irrationality on the part of the agent: even if an agent has excellent evidence that he will decide and act in keeping with

[54] The remaining options may be incomparable or strictly equivalent. While one might allow that choosing in the first case still allows for a substantive application of normative principles, this is far more difficult to maintain in the second. The second is plausibly a case of 'picking', rather than 'choosing' (Ullmann-Margalit and Morgenbesser 1977).

[55] See Levi (1986, pp. 61–4; 1997, pp. 25, 78–80). For further discussion of the problem, see Jeffrey (1977), Shackle (1958, pp. 21–2), Schick (1979,) and Gilboa (1999). Pears also considers such cases (1968, pp. 101–2), in responding to an argument by Cox (1963).

[56] In fact, Levi requires agents not to have any credence whatsoever that they will choose rationally, since he commits to the much stronger claim that agents cannot have any credence in their decisions (see sections 2.2b and 3.1c). With this stronger requirement, the concerns I raise below carry even more force.

what he believes he ought to do, he will not believe he will, or will lose this belief by some unspecified means at the moment of decision. The response is *ad hoc*, requiring a systematic breakdown in our theoretical reasoning capacities to allow for practical deliberation. While weakness of will may explain why some agents never form the belief, it is difficult to see why it should be a standing condition on practical rationality. And note that the problem, in its most pressing form, requires agents to be perfectly rational in other ways. Secondly, this response requires agents to give up on one of the central beliefs that accounts for why normative reasoning is relevant for practical deliberation. If the agent no longer accepts a strong connection between deciding what he ought to do, and deciding what to do, normative deliberation becomes apparently irrelevant to practical decision. Particularly if normative reasons are taken to be essential for practical deliberation, one should not require agents to give up on the belief that they will choose rationally.

Other responders have argued that agents won't recognise their available options prior to the moment of decision. Shackle, for example, ascribes an 'ultimate creative originality' to the agent, the ability to create 'pure visions of states to be desired' (1958, pp. 21–2). Through this ability the agent creates apparent options at the moment of decision that did not exist and were not predictable beforehand. Shackle also seems to identify our freedom in deliberation with this ability (1958, pp. 21–2). But while there may be cases in which we are uncertain of the available options in advance, there are also cases in which the options are clearly demarcated and recognised in advance. Grant's deliberation might be very simple—to take his daughter to the park, or not. Agents still deliberate in such cases. Jeffrey also responds to the problem along similar lines. He claims that an agent may not yet recognise her options and preferences, or may not be certain her beliefs and preferences will not change (1977, pp. 77, 137). So the set of options that are determined by normative principles, given her current beliefs and preferences, do not restrict her available options at the moment of decision.[57] Again, this may account for some cases, particularly when the moment of decision is far off. But as agents approach the moment of decision, their beliefs and preferences will be just as accessible in making the decision than they are prior to decision. We shouldn't require agents to presume their beliefs, desires, or available options will change, or require them to be unknowable in advance of decision.

Instead, I recommend a simpler response—give up the idea that forming a normative belief is always separated, temporally or otherwise, from making a practical decision. Say Grant is committed by a previous standing decision to decide and act as he believes he ought. And say it is by way of this commitment that he is certain he will decide as he ought. In this case it is not surprising if by working out what he ought to do, he thereby becomes decided on a particular option. Given his standing commitment, his making a normative decision on what he ought to do immediately implies his making a decision about what to do. No further deliberation is required. And it is clear why, for Grant, considering normative reasons is essential to practical deliberation—it is the necessary means by which he works out what to do. Consider an analogy. I may decide to act on the result of a coin toss, and buy the feathered hat just in case the coin lands tails. In this case, by the coin landing tails, I thereby become decided on a particular action option. There is no need for me to then deliberate on whether I will act in accord with the coin—I'm already committed to this. This is not to say

[57] Schick is not explicit about which particular beliefs an agent should give up. But he argues that since we are not committed to full self-knowledge, it should not be disturbing if rationality and the logic of decision-making put limits on what we can justifiably believe about ourselves (1979, p. 244). Searle also requires that deliberation only starts where our beliefs and desires leave off, but without directly addressing the problem above (2001, pp. 13−4).

that I can't revise my decision and commitment on seeing the result—one of the main uses of the coin toss being to reveal one's hidden preferences. But this is a case of revisiting a prior decision—nothing in original decision was irrational or inconsistent.

If we offer this last response to the problem, and allow agents to keep their firm beliefs that they will decide rationally, we can acknowledge a strong role for normative reasoning in deliberation. An agent's practical choices may be guided by normative reasons. Provided the agent's belief that she will act in accord with what she takes her reasons to be is sufficiently strong, the impossibility thesis percolates up to the normative level—practical deliberation requires the agent to be uncertain of what she ought do.[58] The normative question of what one ought to do and the practical question of what to do won't always be settled together—one may not be committed to acting as one ought, be sceptical of whether the normative covers all cases, or simply sceptical that one has sufficient information and resources to work out what one really ought to do—but acknowledge that a decision has to be made anyway. But nothing about characterising deliberation in epistemic terms prevents acknowledging a place for normative reasoning in practical deliberation.

The difficult remaining cases are those in which an agent is certain she'll decide or act as she believes she ought, but where she doesn't have a standing commitment to do so. Instead, the agent merely has reliable evidence of these correlations—from observation and induction, or reliable testimony. If such an agent is to practically decide, having decided what she ought to do, one of the other responses that I gave to the Sandy case must be given. But in this case,

[58] We may even account for an agent's reneging on her decision in this way. If an agent revises her belief about what she ought to do, this may lead to uncertainty about what she will do. So she can then re-enter deliberation.

it is also plausible that such an agent does *not* practically deliberate—at best she can find out what she will do, using evidence of what she ought to do.[59] The normative belief gets in the way of her being a practical deliberator. Such cases don't rule out rational deliberation in general. Instead they reveal the connection between the apparent availability of options and gaps in our ordinary evidential beliefs.

### 2.4 Explaining our Apparent Freedom in Deliberation

Having considered some of the upshots of the impossibly thesis, we're in a position to consider how these epistemic requirements on deliberation can explain our apparent freedom in deliberation. To begin, a few notes on the explanandum. I take the apparent freedom to be explained to be an apparent freedom to decide in different ways. I deliberately specify 'apparent freedom' in a very general fashion. This is because I want to be open-minded about how such freedom might be explained. I also allow that this apparent freedom might be manifest in a variety of ways. It may manifest phenomenologically (thought not necessarily as any *distinct* phenomenology), or as a disposition to assent to being (or appearing) free, or more generally as a *belief* that one is (or appears) free. If a belief is involved, it need not be high-level or conceptual; agents need not be able to articulate their belief. I also leave the sense of possibility involved in an agent appearing 'able' to decide generic and unspecified. A similar generic sense is appealed to in principles like 'ought implies can'. Finally, I don't assume the appearance of freedom is either veridical or illusory—we may or may not be free. As I argued in chapter 1, explaining our apparent freedom doesn't directly settle whether we are free.

[59] One might also hold a more minimal epistemic view of deliberation, where practical deliberation is always merely a process of finding out what one will do. I have no objection to such a position, but the epistemic model does not commit one to it. One can still maintain a distinction between theoretical beliefs about what is the case and practical decisions about what is to be the case.

How does the epistemic model explain our apparent freedom in deliberation? It does so by accounting for the openness and unsettledness of deliberation, as well as the evidential relevance of our practical decision for the world. As Jacinta deliberates on whether to stay home or travel on the coast, she is not sure of what she'll decide to do or do. By condition *a*, for any available option she must be unsure of whether she will decide that way. By conditions *c* and *d*, she must be unsure of whether the available option will actually obtain. And by conditions *e* and *f*, she must take several incompatible options to be available. These requirements capture the uncertainty involved in deliberation. As she deliberates, Jacinta's beliefs are genuinely unfixed concerning what she'll do and decide at the conclusion of her deliberation. Note the important role being played by the impossibility thesis: it is by being ignorant of what she'll decide, and what will happen, that there are different options available to her. This accounts for one aspect of the 'openness' involved in deliberation.

But it is not ignorance alone that accounts for an agent's sense of freedom. To distinguish cases of practical deliberation from mere ignorance, I put forward an additional requirement: the agent must be certain that the option will obtain, given that she decides on it (condition *b*). Jacinta must be practically certain that if she decides to stay home, she will stay home. In addition, Jacinta cannot be certain that she'll stay home regardless of her decision (condition *d*). Together, these conditions captured a minimal sense in which we take our decision to be *epistemically relevant* for the state of the world.

The epistemic relevance of the decision for the outcome is crucially what separates practical deliberation from theoretical reasoning. In working out what is the case in general ('theoretical reasoning'), we are also usually ignorant of what is the case. But we can't take the conclusion of theoretical reasoning (a judgment, belief or decision about what is the case) to be a reason for thinking that very same thing is the case. Raamy can't decide crocodiles are dangerous, and use this decision as a reason for thinking crocodiles are dangerous. To do so risks boot-strapping, making any theoretical belief or decision he makes self-reinforcing. In practical deliberation, an agent decision does nothing wrong in taking her decision to be good evidence for what is the case.

It is the combination of *uncertainty* about her decision and the world, yet certainty about the *relevance* of her decision for the world that explains an agent's apparent freedom in deliberation. Different decisions appear as epistemic possibilities for Jacinta. And she takes these decisions to be directly relevant for how the world goes. She seems able to make different decisions because she takes it she might, and takes the state of the world to depend epistemically on her decisions. These conditions explain why an agent's takes various options to be available to her and takes these options to be determined by her decision. So they explain why an agent appears free to decide in different ways in deliberation. Agents like Jacinta can reasonably deliberate on different alternatives when the epistemic conditions are satisfied. And those same epistemic conditions explain why she appears free to decide in different ways.

The conditions also explain why the feeling of freedom fades as deliberation concludes. Before a decision is made, an agent is necessarily uncertain about what she will do. But when an agent actually decides, and thereby becomes certain of her decision and what she will do, the epistemic conditions are no longer satisfied. Her decision space of available options collapses. The selection of one options means the other options are no longer serious possibilities for her. And when there is only one available option, there is nothing left for her to deliberate concerning. The agent may still revisit the decision-problem, by reneging on her decision. She may also still fall prey to weakness of will, and not carry out her decision. But in either case, re-entering the deliberation will require giving up certainty in the outcome.

Some will likely find this epistemic explanation of apparent freedom too deflationary or lightweight to be satisfying. I noted above that the conditions themselves are minimal, and many would prefer characterising deliberation in causal or agential terms. Action theorists and those interested in practical normativity may wish to characterise the options chosen as actions, and also stipulate that the decision is arrived at by a particular means—through considering normative reasons, for example, or in a reason-responsive way. Causal decision theorists may require the decision be taken to cause the option, rather than to merely provide epistemic certainty.[60] In response, let me note some of the features of this epistemic explanation, as well as some of the other ways that causal and agential conditions can still be involved.

Firstly, the epistemic explanation appeals to structural conditions required for deliberation—what background epistemic conditions must be satisfied for deliberation to take place. These conditions were also put forward as constitutive norms, to demarcate a kind of activity that serves an important epistemic function. By appealing to structural and normative conditions, the epistemic explanation focuses on the background conditions required for deliberation to begin and be sustained, and less on how deliberation looks 'from the inside'. Recall, the explanation doesn't attempt to characterise or explain a particular phenomenology, or

[60] Note that causal decision theory is not committed to such a requirement, since the theory is about what it is rational for us to decide on, rather than requirements on deliberation per se.

explicate a particular belief in freedom. I don't intend to rule out more particular explanations being given for aspects of deliberation. But these don't rule out a more high-level epistemic account being explanatory.

Secondly, the epistemic model allows for the possibility that causal and agential notions have become bound up with our concepts of deliberation and decision. This is why I describe the epistemic model as a 'model' or 'characterisation' of deliberation, rather than a full account. But this doesn't mean that the model itself is unexplanatory when we abstract from causal and agential features. With respect to causation, for example, I agree with causal decision theorists that causal concepts are relevant for deliberation; but I wish to come at this connection from the other direction, and explain causal notions via their relevance for practical deliberation, rather than presume causal notions from the start. Similarly, with respect to agency, recall that abstracting from agential notions allows for these to be explained by appeal to deliberation. For example, we might explain how we have a sense of direct control of bodily actions by appealing to the fact that we observe reliable evidential connections between our decisions and the movement of our bodies. In this way, agential notions can become bound up with deliberation, consistent with the epistemic model.

Thirdly, the epistemic model does not attempt to explain *how* the epistemic conditions are in fact satisfied. It may be that when Jacinta takes her deciding on taking a coastal trip to settle that she will, this is because she assumes that her deciding is a *cause* of what she will do or because she is committed, as an agent, to acting as she decides. Both are consistent with the epistemic model characterising deliberation and explaining our apparent freedom. The epistemic model can explain our apparent freedom, and the relevance of causation to deliberation, even if there is feedback between our causal concepts and our concepts of action and deliberation.

I also argued (section 2.3) that normative notions could play an essential role in deliberation, consistent with the epistemic model. Other agential features, such as desires, might also play essential roles. We might require, for example, that decisions be formed in response to desires. But there are advantages to abstracting from particular accounts of normativity, desires and the finer-structure of agency and to allowing that we arrive at our decisions via a variety of means. Doing so means defending the account does not commit us to a particular account of practical normativity. Appealing to broader structural conditions on deliberation means we can relate agency to physics and science—however these agential and normative debates work out.

Finally, note the model does not appeal to counter-causal or metaphysical notions of freedom. I don't require agents to be able to decide in ways not determined by previous states of the world and causes. The epistemic conditions appealed to are perfectly compatible with a law-based understanding of the world and the model puts no pressure on determinism, laws or causal closure. It explains our apparent freedom in a way compatible with our scientific picture of the world. If we're to make sense of causation by appeal to deliberation and agency, this is precisely the kind of account we need.

### 2.5 Relating Practical and Theoretical Reason

I end this chapter by considering what the model tells us about the relation between practical and theoretical reason. Along the way, I'll reject two ways in which one might be tempted to read the epistemic model. I'll argue that while the epistemic *conditions* I've argued for are strictly compatible with relating practical and theoretical reason in various ways, the epistemic *model* takes theoretical reason to limit practical deliberation.

The epistemic requirements above make claims about the relationship between an agent's beliefs and what options she takes to be available in deliberation. Having a certain deliberative state (taking an option to be available) implies being in a particular epistemic state (such as taking the option to be a serious possibility). But the requirements in and of themselves make no claim about direction-of-fit: about whether an agent's deliberative state determines her epistemic state, or whether her epistemic state determines her deliberative state. For example, consider the reliable predictor Sandy. Say Jacinta hears Sandy's prediction of her decision. What might Jacinta do? She might form an appropriate belief based on Sandy's prediction, and no longer deliberate. This is an example of her epistemic state determining her deliberative state. Jacinta's beliefs, particularly beliefs arrived at through evidence, limit and determine the state she is in as a practical deliberator. So theoretical reason limits practical reason. Or Jacinta might mistrust Sandy's prediction precisely *because* she takes it she is deliberating and can decide on different options. This is an example of an agent's deliberative state determining her epistemic state. Under this approach, our apparent freedom to deliberate takes precedence over our beliefs, and we simply won't form the beliefs that would prevent deliberation. In this case, practical reason limits theoretical reason.

Which direction of fit should we expect under the epistemic model? We should take theoretical reason to limit practical reason. If Jacinta acquires evidence of what she will do, she will no longer deliberate on that option. While there may be some situations in which an agent's sense of her own free choice overrides her regular belief-forming practices, and so influences what she takes to be the case, this will not be the norm. This is the approach I will take in the chapters to come. Why take theoretical reasoning to limit practical deliberation? Firstly, this approach allows agents to satisfy an important norm of theoretical reasoning: that changes in belief should be justified on epistemic grounds. If Jacinta has good evidence of Sandy's reliability as a predictor, and forms the belief that she is a reliable predictor, the fact that this belief gets in the way of her deliberating is not *epistemic* grounds to abandon the belief. So we shouldn't require Jacinta to give up the belief. Conversely, taking an agent's epistemic state to limit her deliberative state does not, in and of itself, imply that deliberative and practical norms are violated.[61] Practical norms typically apply to the options an agent takes to be available. Jacinta can only be faulted for choosing badly among the options she takes to be available. Even if she may have a reason (in some sense) to do something she does not take it she's free to do, we wouldn't typically assign non-epistemic blame to her for failing to decide on an option she takes to be unavailable. The fault, if there is one, is an epistemic one of being mistaken about what her available options are.

Secondly, taking epistemic reasoning to limit practical deliberation allows the epistemic model to do interesting explanatory work: the epistemic model can explains features of practical deliberation in terms of our epistemic access to the world. In chapter 5, for example, I'll explain why we deliberate before we decide by appealing to an asymmetry of memory—the fact that we have memories of the past, but nothing like memories of the future. Such explanations are only available if theoretical reason limits practical reason.

[61] There are a number of controversial issues in this area. All I wish to claim is that we are not necessarily lead to violations of practical norms, whereas we are lead to violations of theoretical norms.

For these reasons, I take theoretical reason to limit practical reason, rather than the reverse. But there's a third possibility as well. Perhaps the practical-deliberative and theoretical are two independent *perspectives*, such that epistemic requirements only apply from *within* the practical perspective. Rabinowicz suggests such a view when discussing Levi's work: even if there are limitations on the beliefs an agent adopts during deliberation, these don't apply to the beliefs she adopts in her 'purely cognitive or doxastic capacity', but only to the beliefs she adopts from her practical perspective (2002, p. 92). Agents might be able to switch between these perspectives with ease. And while the two perspectives can inform one another, they are largely independent.[62]

This perspectives approach prevents beliefs formed in ordinary theoretical reasoning from limiting practical deliberation. It shares this in common with the view that practical reason limits theoretical reason. Both these approaches allow agents to simply drop or fail to have beliefs (at least from the practical stance) that get in the way of practical deliberation. So the perspectives view runs into the equivalent problems. Firstly, it countenances belief-forming practices that are in violation of epistemic norms: beliefs may be formed, not formed, or dropped, independently of an agent's epistemic grounds (even if only within the practical perspective). Secondly, taking the perspectives approach means we can no longer explain features of practical deliberation in terms of features of our epistemic relation to the world. In addition, this third alternative faces an additional concern. The requirements I argued for above show how beliefs and deliberation relate. But if these were merely requirements that held from within a deliberative perspective, we'd need to introduce further additional

[62] Elements of this view are also found in Ramsey (1929), Spohn (1977, 2007), Levi (1997, p. 80) and Grice (1971, p. 6). Some of these alternatives are discussed in section 3.1d. Note that taking this third option does not automatically solve the problem above (section 2.3) of how normative beliefs may prevent deliberation, since one can still form beliefs within the practical perspective.

requirements to account for how the perspectives relate. So it seems we would still need to consider further epistemic requirements on deliberation.

Under my preferred approach, theoretical reason limits practical reason. We may still talk of adopting different perspectives towards the world as practical deliberators and theoretical reasoners. But these perspectives aren't so independent that beliefs from one don't carry straightforwardly over to the other. Adopting this approach means that while the epistemic model remains perfectly compatible with physical laws, there is one way in which laws pose a threat to practical deliberation. If our epistemic abilities improved to such a degree that we were able to predict our own decisions and actions, then we would lose the ability to freely deliberate. Nothing in the laws being deterministic implies that we will be able to achieve this—but it leaves open the possibility.[63]

## Conclusion

I've argued for an epistemic model of deliberation through defending a set of epistemic conditions on the beliefs we can have while deliberating. We must be uncertain of our decisions, and yet certain our decisions settle what option obtains. It is because we occupy an epistemic middle-ground, having some certainties, but being ignorant of much else, that deliberation is allowed for. These conditions are justified, given that deliberation serves an epistemic function—deciding is a way of settling what will be. While I haven't characterised deliberation by appealing to its normative function, these requirements don't rule out normative reasons playing an essential role in deliberation. The epistemic model that derives

[63] Some deterministic laws have built-in limitations on what one can know of the system, preventing such predictions, such as Bohmian mechanics. There are also computational arguments that even with deterministic laws, agents cannot predict their own future behaviour (Dennett 1984, p. 112).

from these conditions provides a sufficiently rich characterisation of deliberation to do useful explanatory work. The model explains why agents appear free to decide in different ways as they deliberate. And it also has direct implications for the relation between practical and theoretical reason—practical deliberation is only possible because of pre-existing gaps in our theoretical beliefs about the world.

In the next chapter, I go on to distinguish this model from other competing epistemic accounts of deliberation. I'll argue that competing accounts fail to explain our apparent freedom in deliberation in epistemic terms, despite their aspirations. If we're to account for causation in terms of deliberation, the epistemic model is the place to start.

# Chapter 3

## Competing Conceptions of Agential Freedom

*In a sense my present action is an ultimate and the only ultimate contingency.*
Frank P. Ramsey (1929, p. 146)

As we deliberate on whether to do this or that, various options appear to us as possibilities we are free to decide on. How might we account for this apparent freedom? In the last chapter, I argued that an epistemic model of deliberation can explain an agent's apparent freedom in deliberation. This model characterises deliberation by appealing to the beliefs agents can and can't have while deliberating. I argued, for example, that agents can't reasonably deliberate if they already know what they're going to do. Yet agents must take their decisions to epistemically settle what they will in fact do. This ignorance-based model is not the only epistemic model on the market. Competing epistemic models have been used to characterise deliberation, to explain our apparent freedom, to explain the direction of causation and to directly argue against the ignorance conditions I've argued for. David E. Velleman, James M. Joyce, Jenann Ismael and Huw Price all defend competing epistemic accounts. Isaac Levi and Tomis Kapitan defend variants of my ignorance-based account. If any of these competing accounts are successful, the epistemic model is not a good place to start for developing a deliberative account of causation.

The work of this chapter is to show that competing epistemic models aren't a threat to my ignorance-based model of deliberation. I won't consider competing models of deliberation in general—just epistemic ones. As I argued in the last chapter, epistemic models are good candidates for explaining causation without relying on causal or agential features of deliberation.

One competing type of account, defended by Velleman, Joyce and Ismael, attempts to explain our freedom in deliberation by appealing to the fact that we can form certain beliefs 'unconstrained' by evidence. These beliefs constitute our decisions. Velleman (1985, 1989a, 1989b) takes decisions to be beliefs about actions; Joyce (2002) and Ismael (2007, 2012) take decisions to be beliefs about decisions. We have epistemic freedom because we can form these beliefs unconstrained by evidence. However, it is controversial to label these models as 'competing'. Huw Price (2012) has argued that ignorance accounts like my own have their source in these unconstraint accounts. If so, the epistemic model wouldn't be the right foundation for a deliberative account of causation (even if it might explain our apparent freedom). I will argue that Price is mistaken, and that these models are importantly distinct.

I'll also argue that unconstraint accounts face serious problems if used to explain apparent freedom. Firstly, agents turn out to have epistemic freedom even over beliefs formed in response to evidence. Secondly, I'll argue that these accounts presume, incorrectly, that agents are justified in *forming* these decision-beliefs (not just justified in the beliefs once they're formed). But agents aren't so justified. So such justification can't explain apparent freedom. Underlying these problems, I'll argue that these accounts turn out to rely on a further apparent ability to form different beliefs. But once we introduce such an ability, the appeal to epistemic freedom becomes redundant—there are simpler ways to explain apparent freedom. I'll illustrate this point by appealing to a related account of deliberation by the German Idealist J. G. Fichte. Altogether, unconstraint accounts turn out not be successful rivals to the ignorance account.

I begin in section 1, by presenting the competing models, and narrowing the field to those that are the most compelling alternatives to ignorance accounts—unconstraint accounts. In section 2, I argue that ignorance and unconstraint accounts are indeed distinct, and, furthermore, should not be combined. In section 3, I raise two serious problems for unconstraint accounts, which ignorance accounts avoid. Finally, in section 4, I diagnose what I take to be the underlying problem—unconstraint accounts rely on a further apparent freedom to believe in different ways of a type that makes the epistemic explanation of apparent freedom redundant.

### 3.1 Epistemic Accounts of Deliberation

As Tamsin deliberates about whether she will head to bed or stay up reading, these options appear to her as possibilities she is free to decide on. How might we account for this apparent freedom? We might appeal to beliefs—ordinary beliefs about the world, bodily motions, and mental states (rather than beliefs in metaphysical freedom or normative facts). The core idea of a number of accounts is that an agent's beliefs while deliberating do not compel her to decide any particular way. So there is a kind of 'epistemic gap' between her beliefs or evidence and her decision. This epistemic gap gives her 'epistemic freedom' regarding her decision—a term of art from Velleman (1989a). This epistemic freedom might account for why agents seem free to decide in different ways. In this chapter, I'll interpret these various epistemic accounts, including my own, as giving stipulative definitions of different *types* of epistemic freedom—rather than as offering competing analyses of a single phenomenon. Velleman, Joyce, Ismael, Levi, Kapitan, Price and I all give accounts of deliberation that include an epistemic gap. Velleman, Ismael, Kapitan and I also explicitly use our accounts to explain our apparent freedom in deliberation. I'll argue that the competitors to my own account don't adequately explain our apparent freedom.

Before I begin, a few preliminaries. The accounts I consider were developed in different areas of philosophy. To compare them, I have been flexible regarding what kind of apparent freedom they might explain. They might explain a phenomenal feeling of freedom (Velleman) or openness (Ismael, Velleman) or a belief in freedom or openness (Kapitan). Again, I don't presume the appearance of freedom is veridical. I also consider accounts of epistemic freedom that weren't developed to explain apparent freedom at all. Joyce and Levi primarily use epistemic freedom to delimit what options are rationally available to agents. But we should expect a close connection between what options are rationally available, and what options we appear free to choose. And my main concern is whether these accounts *can* be used to explain our apparent freedom in deliberation, on the way to giving a deliberative account of causation—not whether they have been used to do so.

As in chapter 2, I follow the majority of accounts in discussing 'decision' rather than 'intention'. While Velleman uses the term 'intention', he characterises intentions as what we're 'decided' on, and distinguishes them from the goals with which we act (1989a, pp. 112–3). I'll discuss both the act or process of deciding (forming a decision) and the product of that act (the decision). And I'll assume decisions both conclude deliberation and imply beliefs about what one will do. This is common ground between the accounts—for discussion, see section 2.2a, as well as Hampshire and Hart (1958), Pears (1968), Grice

(1971) and Velleman (1989a, p 113).[64] Some accounts take agents to decide on actions; they decide on what they *will do* (Velleman, Kapitan). Other accounts take agents to decide on propositions; they decide *that* something is so (Levi, Joyce, Ismael). As explained in the last chapter (section 2.1), while I prefer the proposition formulation, none of my arguments depend on how options are characterised. For simplicity, and ease of quoting, I will switch freely between the two formulations.

### 3.1a Freedom to Form Self-fulfilling Beliefs

How might we explain our apparent freedom in deliberation? Velleman (1985; 1989a; 1989b) claims agents appear free because they're licensed to form a variety of *self-fulfilling* beliefs about what they'll do. These are beliefs that cause the very states of affairs they represent—such as when my belief that I'll talk confidently brings it about that I do. Velleman argues that certain self-fulfilling beliefs constitute decisions. Say Tamsin is deliberating about whether to go to bed or stay up reading. She is somewhat motivated to go to bed, and somewhat motivated to stay up. Velleman claims that, as an agent, Tamsin has a standing desire to know *what* she's doing as she begins to do it. This motivates her to act as she expects. So if she believes she'll stay up, this additional motivation may be enough to tip the balance and cause her to stay up. And similarly if she believes she'll go to bed. If so, her beliefs are self-fulfilling.

For beliefs to count as decisions, the agent has to also *know* they are self-fulfilling. Decisions are self-fulfilling beliefs that 'represent themselves' as such (1989a, p. 98). Once Tamsin has

[64] This assumption is slightly weaker than what is captured by condition *b* of the epistemic model. Condition *b* requires agents to be practically certain that if they decide on the option, the option obtains—which means decisions will typically imply certainties. The shared assumption here only requires decisions to imply plain beliefs.

formed a decision-belief, she has evidence of its truth—she has the belief, and knows it's self-fulfilling. And so, Velleman argues, her beliefs are still responsible for being true, still governed by appropriate norms, and still count as beliefs (1989a, pp. 56–64, 127–9). Velleman claims, moreover, that Tamsin can be 'licensed' to form a variety of self-fulfilling beliefs, safe in the knowledge that what she believes will turn out to be true. She's *unconstrained* by evidence of what she'll decide or do—she has 'epistemic freedom' (1989a, ch. 5; 1989b; 2000, pp. 22–6, ch. 2). This is 'the freedom to affirm any one of several incompatible propositions without risk of being wrong' (1989b, p. 73). She has this epistemic freedom whenever she knows, for at least two beliefs that are incompatible, that either would be true if formed. Because she is licensed to form beliefs as she likes in these cases, no wonder she appears free to decide in different ways: 'we mistake the license to affirm any one of various things about what we'll do for the (metaphysical) possibility that we might do any one of those things' (1989b, p. 74).[65]

Evidential *unconstraint* plays a central role in Velleman's account. What makes an agent 'feel free…[is her] freedom from the evidence' (1989b, p. 77). Tamsin is not constrained by evidence about what she will decide or do. Even if she has evidence that she won't go to bed, or won't form the belief, she can still know that the belief *would be* self-fulfilling, were she to form it. This gives her epistemic freedom. She is 'licensed' to form the belief, even if the totality of her evidence is conclusive and guarantees the belief's content is false (1989a, pp. 149–153, 165–6, 1989b, pp. 78–80). So freedom is not a matter of ignorance. Adapting an example from Anscombe, Velleman (1998b, p. 79) argues that a doctor is licensed to

[65] Related accounts include Harman (1976) and Joyce (2007, pp. 556–61). Harman takes decisions to merely 'involve' beliefs, and doesn't use beliefs to explain apparent freedom. Joyce generalises Velleman's account to include common cause structures.

make various predictions of where a nurse will take a patient, even when he knows where the patient will go:

> The evidence…contains one component that licenses the doctor to contradict what all of the evidence, taken together, conclusively proves. This component of evidence shows that the doctor would be correct in predicting whatever he likes, within reason, irrespective of what the totality of evidence demonstrates is bound to occur…The central evidence shows the doctor that even if the totality of evidence guarantees one outcome, he would still be correct in predicting others. It shows, in short, that he is epistemically free.

There are some ways in which evidence limits epistemic freedom (1989a, pp. 158–9; 1989b, n. 5). Tamsin must know her belief would be self-fulfilling. If she has evidence that she's been glued to her armchair, for example, or her motivation for going to bed is very weak, she won't be epistemically free to decide to go to bed. But if she knows the relevant counterfactuals are true, she is licensed to form different incompatible beliefs.[66]

If Velleman's account succeeds, we have an explanation of how we usually know what we're doing as we begin to do it—forming a decision simply is forming a justified belief about what we're going to do. And we seem to explain our freedom in practical deliberation in terms of something much more familiar—the kind of justification we have in theoretical reasoning. If Tamsin really is licensed to form different beliefs about what she's going to do, this might explain how she appears free. While Velleman's explanation relies on Tamsin having a desire to know what she's doing, perhaps such a desire isn't so strange. Our lives wouldn't go well if we often didn't know what we were doing. And this desire need not be conscious or person-level.

[66] Velleman acknowledges that actions that we're highly motivated towards can't be decided on, since the beliefs would not be self-fulfilling (1989a, p. 159). This is a concern for Velleman's account, but one I won't press.

To summarise, Velleman's account has the following important features. It identifies decisions as beliefs. It takes these beliefs to be justified once formed. And it uses an agent's license to form these beliefs to explain apparent freedom. The special license agents have in decision contexts means they aren't constrained by evidence of what they'll do. This epistemic freedom looks like just the thing to explain apparent freedom—it is a freedom agents only have while deliberating that does not rely on metaphysical notions of freedom.

### 3.1b Freedom to Form Self-constituting Beliefs

Joyce (2002, pp. 94–8; 2007, pp. 556–61) and Ismael (2007; 2012) defend variants of Velleman's account. They agree with Velleman that decisions are beliefs, and that agents form these beliefs unconstrained by evidence. But, rather than taking decisions to be beliefs in actions, they take them to be beliefs in the very decisions themselves. The belief in the decision *is* the decision. Tamsin's deciding to go to bed, for example, *is* her forming the belief that she decides to go to bed. 'The volition stated in the first person at the conclusion of a piece of deliberative reasoning ("I will that so and so") is at once volition and belief about my volition' (Ismael 2012, p. 154). In Joyce's terms, '*believing* that one has decided to do *A* can, under the right conditions, *be* a decision to do *A*. The "right" conditions are just those of deliberation' (2002, p. 97). The belief does not cause the states of affairs it represents—it constitutes the state of affairs.[67] The belief in the decision is *self-constituting*. While Joyce and Ismael call these beliefs 'self-fulfilling', I'll reserve that term for beliefs that are *casually* self-fulfilling.

[67] Does this identification lead to an infinite regress of beliefs in decision? While Joyce and Ismael don't discuss this, presumably the decision gets its content through its causal connections to action, and so a regress is avoided. These reflexive beliefs are still odd, but I won't take that as a reason to reject them.

Ismael (2007, p. 5; 2012, p. 155) and Joyce (2002, p. 98) liken these self-constituting beliefs to speech acts called *performatives*. When Ben asserts 'I hereby promise…', his act of asserting constitutes his promising. In asserting, Ben is unconstrained by prior evidence of what he will promise. Joyce and Ismael claim that because beliefs in decisions are similarly self-constituting, we form them unconstrained by evidence. 'The decision-maker is 'never "hemmed in" by her evidence' (Joyce 2002, p. 98). In Ismael's words (2012, p. 154):

> evidence is irrelevant *because* I cannot be wrong. My beliefs about my own pending decisions (are) epistemically unconstrained by any information I might have…any such information is automatically overridden by the decision process itself, and (hence) it can't *constrain* its development.

An agent can form whatever belief she likes, safe in the knowledge that it will turn out to be true. There are still some constraints. For a belief in a decision to be a decision, the belief must be formed 'during' deliberation on how to act (Joyce 2002, p. 97) or at the 'conclusion' of such deliberation (Ismael 2012, p. 154). And, at least for Joyce, the agent must believe that the decision will cause the act (2002, p. 98). But, within these constraints, agents can form a variety of incompatible beliefs, unconstrained by evidence—they have 'epistemic freedom' (Joyce 2002, p. 96), a connection Ismael also notes (2012, p. 156).

Joyce and Ismael's unconstraint account of epistemic freedom seems to deliver the same important benefits as Velleman's. It doesn't rely on metaphysical notions of freedom. And it can explain our apparent freedom in deliberation—we appear free because we can form a variety of incompatible beliefs, unconstrained by evidence.

### 3.1c Competing Ignorance Accounts

In chapter 2, I argued for an ignorance-based epistemic model of deliberation. This model characterised deliberation using six necessary conditions on deliberation. An agent takes an option **a** to be an available option in deliberation only if:

(a) her deciding on **a** is a serious possibility for her;

(b) she is practically certain that if she decides on **a**, then **a**;

(c) **a** is a serious possibility for her;

(d) it is a serious possibility for her that if she does not decide on **a**, **a** does not obtain.

(e) she takes several options to be available;

(f) she takes her apparently-available options to be incompatible.

I argued that these conditions are individually necessary for deliberation, and jointly characterise deliberation. I also argued that they explain an agent's apparent freedom in deliberation. An agent appears free to decide in different ways as she deliberates because her beliefs leave open how she will decide, yet she takes her decision to settle what option obtains.

In keeping with this model, I take an ignorance variety of epistemic freedom to be characterised by the following requirement: an agent must be ignorant of what she will decide and do if she is to be epistemically free. This ignorance-based gap between her beliefs and decisions may then explain her apparent freedom in deliberation. If so, we have an ignorance account of apparent freedom. Unlike unconstraint accounts, ignorance accounts need not take decisions to *be* beliefs. Decisions need only imply beliefs about what one is going to do.

Levi and Kapitan both give accounts of decision-making that appeal to ignorance conditions. Kapitan argues for the following two conditions on an agent taking an action to be an open option (1986, p. 241; see also 1984; 1989; 1990):

> an agent presumes that his ø-ing is an open course of action for him *if and only if* (i) he presumes that he would ø *if and only if* he were to choose to ø, and (ii) he presumes that if S is any set of his beliefs then his ø ing is contingent relative to S.

According to the second condition, an agent will only take an option to be available if she *presumes* she is ignorant of what she will do, taking it that none of her beliefs either individually, or as a set, or a subset, settle what she will do. Typically an agent will presume this because she is actually ignorant, and does not have beliefs that settle what she will do. Kapitan includes a third condition as well: an agent deliberates between a set of options only if she presumes the options are not all conjointly realizable (1986, p. 243). Kapitan's conditions are similar to my own. I noted some of the minor differences between our conditions in the last chapter, section 2.2a.

There are more significant differences between Levi's ignorance account and my own. Levi and I agree that an agent can only recognise an option as available if it is a 'serious possibility' for her—a possibility not ruled out of consideration by her beliefs (1986, ch. 4; 1997, ch. 2 and 4; 2000, pp. 393–6). This requirement means that an agent must be ignorant during deliberation of what option she will choose, for all but trivial forms of deciding. But ignorance conditions come in various strengths. Levi argues that agents can't properly have beliefs of *any* degree in their decisions while deliberating. Tamsin, for example, can't believe to degree 50% that she will stay up reading, while she deliberates on whether to. Deliberation 'crowds out' prediction of any degree (1997, p. 32). I only argue that agents can't properly have beliefs that give them *practical certainty* of what they will do. I allow that Tamsin can deliberate while having degrees of beliefs in her impending decisions.

Why does Levi think beliefs of any degree are ruled out? His first argument (1997, pp. 32, 76–7) draws on Spohn (1977, pp. 114–5). Levi argues that agents need not assign credal probabilities to options in decision-making, and so as theorists we should not assign them either. What Levi seems to have in mind is the following. Tamsin doesn't need to have any degree of belief about whether she will take her umbrella in order to appropriately deliberate about whether to. She needs beliefs about what will happen *if* she decides to take her umbrella (or not). But beliefs about whether she's likely to don't play a role in her decision-making.[68] However, this argument is not enough to support the claim that agents *can't* properly have degrees of beliefs while deliberating. Even if such beliefs play (or should play) no role in that particular deliberation, they may play important roles in other deliberations the agent is engaged in. For example, Tamsin may also be deliberating on which bag she should pack, at the same time as she deliberates on whether to take her umbrella. If she needs to pack her bag before she decides whether to take her umbrella, it may be useful for her to know how likely it is that she'll take her umbrella, in order to decide which bag to pack.

[68] Levi might also argue that agents *should not* let their beliefs about impending actions play any role in decision-making—even if they sometimes do.

Levi (1997, pp. 32, 76–7) gives a second argument against agents having degrees of beliefs in their impending decision, again drawing on Spohn (1977, pp. 114–5). Levi argues that if we measure degrees of belief by the usual means, offering bets, agents turn out not to have degrees of belief (except 1 and 0) in their impending actions. This is because in decision contexts an agent's willingness to accept bets is no longer sensitive to the odds at which those bets are offered. For example, say we offer Tamsin a bet that gives her $100 if she takes her umbrella and nothing otherwise. If taking her umbrella were out of her control, and Tamsin is 50% certain she would take her umbrella, she should be willing to pay up to $50 for the bet. She'd take herself to be likely to make money off the bet rather than lose money. But she shouldn't be willing to pay $60 if she were only 50% certain. She'd take herself to be likely to lose money. Knowing at what odds Tamsin is willing to accept the bet can then tell us how likely she thinks it is that she'll take her umbrella. However, typically whether Tamsin takes her umbrella *is* under her control. And so provided the money gained matters more to her than the inconvenience or other costs involved in taking her umbrella, Tamsin should be willing to pay up to $100 to take the bet. She should bet, and then simply decide to take her umbrella in order to win the money. This behaviour, by the usual calculus, corresponds to her having a credence 1 (being *certain*) that she'll take her umbrella and having a credence 0 that she won't.[69] So if we assign credences to agents based on the odds at which agents accept bets (and the stakes are sufficiently high), we will be led to assign agents credences in their impending actions of either 1 or 0. And credences of 0 and 1, for the most part, rule out these options as available, given the epistemic conditions Levi and I argue for—agents must be ignorant if they are to non-trivially deliberate and decide.

[69] Note that if Tamsin is offered a bet on *not* taking her umbrella, the situation is reversed. So the credences we assign are not well-behaved in other ways as well.

However, there are a number of problems with this argument. Note that these problems arise independently of whether the model is used to give a deliberative account of causation. Firstly, at best Levi's argument shows that sufficiently-high-stake bets prevent betting behaviour revealing degrees of belief. It says nothing about low-stakes betting. Offer Tamsin a bet that gives her \$1 for taking her umbrella (and nothing otherwise), and the possibility of winning the \$1 will not have much effect (if any) on her willingness to take her umbrella. Only if she is more than 50% certain she will take her umbrella should she pay more than 50c for this bet. If so, her disposition to accept the bet under various odds will (roughly) reveal her degrees of belief. As the stakes diminish, betting dispositions more closely track degrees of belief. Secondly, even if betting behaviour turned out to be entirely inappropriate for tracking degrees of belief in impending decisions, this would not imply that agents lacked such beliefs—just that they are hard to measure. For related arguments and similar conclusions, see Joyce (2002, pp. 81–7) and Rabinowicz (2002).

My responses to both Levi's arguments exploit the fact that agent can be properly involved in several deliberations at once. Tamsin can deliberate on which bag to pack, or whether to accept a bet, while still deliberating on whether to take her umbrella. The decision she makes about the bag or the bet *can* affect what decision she eventually makes about the umbrella. But in many such cases, it will still be useful for her to predict whether she'll take her umbrella, even while deliberating on whether to. One could block my responses to Levi by arguing that agents can't simultaneously engage in several deliberations at once, or that, as theorists, we should always relativise an agent's belief set to a particular deliberation, and take such belief-sets to be relatively independent of one another. But we would need powerful reasons to make these moves. We are often required to engage in multiple deliberations, particularly when our deliberations are long and complex. And we often do use the same beliefs in multiple deliberations. Typically we would want a model of deliberation to approximate what agents actually do.[70] Because I reject Levi's arguments for ruling out degrees of belief, the ignorance account I suggest is more modest, requiring only that deliberating agents aren't practically certain of how they will decide.

### 3.1d Perspectival Accounts

Huw Price has proposed a number of accounts of deliberation that are epistemic in a broad sense. Under these accounts, taking up the *perspective* of a deliberator implies having, or lacking, certain beliefs. These accounts might seem to explain our apparent freedom in epistemic terms. Under one of Price's proposals, taking up the perspective of a deliberator implies viewing one's impending actions as having no causal history, or a causal history independent of other relevant causal mechanisms in the system: 'agents see their own actions as "uncaused," at least in the midst of deliberation about those same actions' (2012, p. 536). 'To introduce the agent is in effect to assume an independent causal history to the event *A*' (1991, p. 169), or 'a history that would ultimately originate in one's own decision, freely made' (Price and Menzies 1993, p. 191). Price effectively suggests a kind of Kantian perspectival incompatibilism. In taking up the perspective of a deliberator, I assume my decisions or actions are uncaused causes, that is, causally independent of (past) features of my psychology and environment, but causes for subsequent (future) worldly events. Price traces this view back to Ramsey: 'From the situation when we are deliberating seems to…arise the general difference of cause and effect' (Ramsey 1978, p. 146; quoted Price

[70] Models of decision theory also idealise in various ways. But the more they do, the less convincing they are as arguments against actual agents having degrees of belief in their impending decisions, and the less suitable they are for explaining our apparent freedom in deliberation.

1993, p. 254—see also ibid. pp. 260–265; 1992, p. 518).[71] Taking my decision or action to be an uncaused cause might seem to imply taking it to be evidentially severed from past events. In Ramsey's words,

> …any possible present volition of ours is (for us) irrelevant to any past event. To another (or to ourselves in the future) it can serve as a sign of the past, but to us now what we do affects only the probability of the future…In a sense my present action is an ultimate and the only ultimate contingency.
>
> (1978 [1929], p. 146)

There's an open question of whether such a view accurately depicts the phenomenology or the perspective of deliberating agents. But there's a deeper problem. If such an account is to be used to deliver a deliberative account of causation, as Price and I intend, it presumes too much in the way of causation. The problem is easiest to see in the case of the temporal asymmetry of causation. Both Price and I aim to explain why causes come before their effects by appealing to deliberation. Under Price's above proposal, being a deliberator requires seeing my impending acts or decisions as evidentially severed from their usual causes (while still being relevant for their effects). But if we ask why my decisions are evidentially severed from *past* states, and not *future* states, we have to import the temporal asymmetry of causation—that the causes of my action lie in the past, and its effects lie in the future. We make no progress explaining the temporal asymmetry of causation. Price could simply stipulate that agents take their decisions to be evidentially severed from past states (as Ramsey does). But this would be to import a temporal asymmetry equally as mysterious as the causal asymmetry we are trying to explain.

[71] Similar ideas are also appealed to by Hitchcock (1996) and Pearl (2000, pp. 108–109). I consider Pearl's view in the next chapter (section 4.2b). For further discussion of the 'Ramsey thesis', see Ahmed (2014, ch. 8).

A similar problem occurs with a related account offered by Price. Price claims that an agent cannot properly take her free acts or decisions to be evidence for their usual causes, while deliberating about those very acts. Why? Because an agent should realise that her deliberation will be an 'alternative causal explanation' for why she decides as she does (1991, p. 166), one which undermines any evidential correlation that otherwise would obtain between her decision and its causes. Price appeals to this claim in defence of evidential decision theory (1986, 1991). And he goes on to suggest that it can be used to explain Ramsey's thesis in 'purely evidential' terms, such that there is 'simply no asymmetric modal notion left unaccounted for' (1993, p. 261; see also 2007, pp. 281–2; Price and Weslake 2009, p. 431).

But unfortunately the same concerns with presuming causation and causal asymmetry apply. Price's argument for undermining relies crucially on the idea that an agent's deliberation provides a competing *causal explanation* for why she decides as she does. Unless we or the agent makes causal assumptions, neither of us have reason to expect such undermining. And if we ask why evidential relations to the past and not the future are undermined, the best we can do is simply rely on the fact that causes come *before* their effects—so that the competing causal explanation provided by deliberation undermines only correlations to the past. There are other arguments that might be used to support the Ramsey thesis and related perspectival accounts. But the suspicion is they will similarly rely too much on causation and causal asymmetry. For now, I put these perspectivalist approaches aside and move on to consider the relation between the two most promising types of epistemic account—ignorance and unconstraint accounts.

### 3.2 Similar, yet Different

Unconstraint and ignorance types of epistemic freedom share features in common. Both include an epistemic gap between an agent's evidence or beliefs *while* deliberating and what she believes *as a result of* decision. An agent's evidence or beliefs while deliberating do not compel her to make any particular decision, or to form any particular belief about what she will decide or do. It is for this reason that these types might seem equally suitable for explaining why agents appear free. In this section, I argue that these accounts are importantly distinct. Neither is equivalent to the other, and, moreover, they should not be combined.

Price (2012, pp. 528–9) argues that ignorance and unconstraint accounts are ultimately different expressions of the same underlying idea. He takes these accounts to enforce the same kind of epistemic gap between evidence and decision—and then uses this gap in giving his own agent-based account of causation. Price begins with Joyce's 2007 account, which characterises decisions as causally self-fulfilling beliefs. Immediately following this he suggests, 'An alternative way to put this thought…is to say that there is an important sense in which, as she deliberates, an agent simply *does not have* knowledge, beliefs, or credences about the action in question' (2012, p. 529, his emphasis). This is an ignorance account—Price even quotes Rabinowicz's summary (2002) of Levi. Finally, Price claims we get to both these views by considering the 'epistemic authority' agents have with respect to their own actions—their ability to trump predictive knowledge (2012, p. 529). He takes the source of this authority to be given by Ismael's account of decisions as self-constituting beliefs.

Price appeals to all three types of accounts of epistemic freedom—a causal account (Joyce 2007), a self-constituting account (Ismael 2007), and an ignorance account (Levi, via

Rabinowicz 2002). What he seems to have in mind is the following. Under unconstraint accounts, even though an agent might seem to have evidence while deliberating about what she'll decide, she shouldn't take this to be evidence. Why? Because she knows she is free to contravene it. Like Dummett's chief (1964), she is free to contradict or 'bilk' her evidence by deciding otherwise (Price 2012, p. 529). Her epistemic authority over her current decision leads to ignorance elsewhere. Ignorance accounts, like Levi's, turn out to have their source in unconstraint accounts, since it is unconstraint accounts that ultimately explain an agent's epistemic authority.[72]

But unconstraint and ignorance types are importantly distinct. They imply different kinds of epistemic gaps. Under ignorance accounts, an agent does not *have* beliefs or evidence that settle how she will decide—this is the epistemic gap between her evidence and what she decides, and why she appears free. Under unconstraint accounts, an agent is epistemically *unconstrained* by her evidence, regardless of what evidence she has. This is the epistemic gap, and why she appears free. Unconstraint accounts hold that the evidence an agent has while deliberating is *irrelevant* to her apparent freedom. Tamsin can be reliably informed she'll go to bed, believe she'll go to bed, and yet sensibly deliberate about whether to go to bed. Under an ignorance account (as I argued in chapter 2, section 2.5) she cannot. Only under an ignorance account does prediction prevent deliberation.

Velleman (1989a, pp. 151–3; 1989b, pp. 78–80) and Joyce (2002, p. 94) in fact use unconstraint accounts to argue *against* ignorance conditions like Levi's. They're not mistaken

[72] Some of Levi's arguments do suggest ignorance arises because of our freedom in deliberation (1997, pp. 76–7). But Levi does not appeal to unconstraint or epistemic authority. And as noted in regards to betting (section 3.1c), Levi's arguments for this point aren't strong.

to do so. Unconstraint accounts do not imply ignorance. If an agent's apparent freedom is explained independently of her beliefs about what she will do, her apparent freedom gives her no reason to forgo these beliefs. What explains her apparent freedom is knowing her beliefs *would be* true if formed. This knowledge is independent or her beliefs about what she *will* do. If having these beliefs does not get in the way of her apparent freedom, the fact that she is deliberating gives her no reason to forgo them.

Why might it have seemed that ignorance accounts had their source in unconstraint accounts? Both Levi and Price argue that an agent's freedom in deliberation implies complete uncertainty of what they will decide and do—in Levi's case, via a betting argument (section 3.1c), in Price's case, by appeal to an agent's epistemic authority over her decisions—of a kind taken to be implicit in Dummett (1964). Price, Dummett and Levi all rely on the idea that no matter what her evidence, an agent can decide how she likes. This makes her beliefs in her impending decisions difficult to track and might make it seem as if she has, or should have, no beliefs in her impending decisions or actions. But claiming an agent has the *ability* to decide in ways not constrained by her evidence is quite different from saying she is *justified* in forming beliefs unconstrained by her evidence. Neither Levi nor Dummett are interested in an agent's justification. Nor are they directly interested in what explains an agent's ability (or apparent ability) to decide in different ways. It is only unconstraint accounts that are taken to explain an agent's (apparent) ability in terms of her justification. Price would be wrong to assume that Levi and Dummett appeal to an agent's justification. However, I will later argue that unconstraint accounts *do* presuppose an agent's (apparent) ability to decide in different ways (section 3.4). So in a sense it's not surprising that Price would appeal to unconstraint accounts to derive ignorance accounts. But what

Price is ultimately relying on is an agent's (apparent) freedom to believe in different ways—not the fact that agents are justified in forming beliefs unconstrained by evidence.[73]

Finally, not only are ignorance conditions not implied by unconstraint accounts, but adding them on top of unconstraint accounts leads to trouble. Unconstraint accounts begin with the idea that an agent's evidence of what she'll decide does not constrain her deliberation. Adding in ignorance conditions would mean that the fact that an agent is deliberating changes the import of her evidence. Even if Tamsin has been reliably told that she will decide to go to bed, the fact that is unconstrained by her evidence would mean she should no longer treat it as evidence. But the fact that an agent is unconstrained by her evidence does not change its import. Her evidence remains just as strong. Ignorance conditions do not derive from unconstraint accounts, and should not be included as further additions.

If we are going to combine ignorance and unconstraint accounts, a more plausible alternative is to take ignorance conditions to be prior constraints on deliberation. If making a decision just *is* forming a belief, this plausibly requires the agent not to already have the relevant belief. If so, ignorance should be a prior condition on deliberation, under unconstraint accounts. This kind of hybrid view is different from a 'pure' ignorance account of the kind I defended in the last chapter. Pure ignorance accounts don't appeal to an agent's justification to form beliefs unconstrained by evidence to explain her freedom in deliberation. Hybrid accounts do.[74] I argue below, however, that unconstraint *doesn't* explain our freedom to

[73] However, even if agents have a primitive (apparent) ability to decide in different ways, this still doesn't imply they are ignorant of what they will decide. This was my argument against Levi in section 3.1c.

[74] Hybrid accounts are similar to ignorance accounts that identify decisions as beliefs. As far as I know, such accounts have never been explicitly advocated—although see Hampshire and Hart (1958) and Pears (1968).

decide. What this means is that even if we adopt a hybrid account, it is ignorance that still does the work of explaining our apparent freedom in deliberation.

### 3.3 The Dangers of Unconstraint

I'll now argue that unconstraint accounts face two serious problems when used to explain apparent freedom. Firstly, agents turn out to have epistemic freedom even over beliefs formed on the basis of evidence. Secondly, agents are *not* justified in forming beliefs unconstrained by evidence—and so such justification can't explain apparent freedom.

### 3.3a Freedom over Evidence-based Beliefs

Velleman, Joyce, and Ismael identify decisions as self-fulfilling or self-constituting beliefs. Our license or justification to form these beliefs might explain our apparent freedom. But a concern immediately arises: beliefs formed on the basis of evidence can also be self-fulfilling or self-constituting. So we can have epistemic freedom even over evidence-based beliefs—a problem for explaining our apparent freedom over decisions.

I'll begin by raising this objection against Velleman's account. Say Tamsin is deliberating over whether she will go to bed. Her partner tells her she will, and, based on this evidence, she forms the belief that she will. Velleman claims Tamsin has a standing desire to do what she expects. This desire gives her additional motivation to go to bed, if she believes she will. Stipulate that this additional motivation causes her to go to bed—without it, she would have stayed up. In this case, her belief formed based on evidence is self-fulfilling. Tamsin may know this belief is self-fulfilling, in so far as she knows her decision-like beliefs are self-fulfilling. There may be a second belief (say a belief that she'll take a bath) that satisfies these same conditions. Under Velleman's account, because these incompatible beliefs are self-fulfilling and represent themselves as such, Tamsin has epistemic freedom over them, and will experience apparent freedom. Yet these beliefs are based on evidence.

If Tamsin has epistemic freedom over evidence-based beliefs, Velleman's account faces a dilemma. It either identifies the wrong beliefs as decisions, or it grants us epistemic freedom over beliefs that aren't decisions—so it can't explain the apparent freedom we experience particularly over decisions. Tamsin's evidence-based belief may need to satisfy various other conditions to count as a decision. She may have to knowingly form it based on evidence of her desires. Perhaps she does. If so, Velleman's account wrongly identifies her evidence-based belief as a decision.[75] Alternatively, Tamsin's evidence-based belief might not count as a decision. But then Tamsin has epistemic freedom over beliefs that aren't decisions. The overall concern is that Tamsin has epistemic freedom over evidence-based beliefs—whether or not they count as decisions.

The problem arises not because Velleman identifies decisions as beliefs, but because he appeals to our *license* to form self-fulfilling beliefs to explain apparent freedom. This kind of license applies equally to evidence-based beliefs. So it does not explain why we have apparent freedom only over beliefs formed *not* based of evidence. Velleman cannot then avoid the problem by taking beliefs to be mere component parts of decisions. The same problem over license would apply. Ignorance accounts of apparent freedom avoid this problem, because they don't appeal to such license to explain apparent freedom.

[75] Bratman (1991, pp. 123–5) discusses a similar concern with regard to known side-effects.

This same objection applies to Joyce's and Ismael's accounts, at least under one interpretation. Say Tamsin is deliberating about whether to go to bed and her partner tells her that she will decide to. Believing this prediction, she forms the belief that this is what she decides. As far as her beliefs in decision generally cause actions, this one will also cause her to act. Because it is a belief in decision formed during deliberation that causes her act, it counts as her decision. We can again stipulate that if she hadn't heard the prediction, she would have decided to stay up. A second incompatible belief (say a belief that she decides to take a bath) can satisfy these same conditions. So Tamsin has epistemic freedom over evidence-based beliefs.

But perhaps this is the wrong way to interpret Joyce and Ismael. Perhaps they don't intend that any belief in decision formed *during the temporal interval* of deliberation counts as a decision. Perhaps they intend to restrict to beliefs formed on the *basis of* deliberation—such as by an agent judging that the option is among her best. Such an interpretation is suggested by Joyce's appeal (2002, n. 41) to Skyrms' deliberational dynamics (1990), which considers how an agent's beliefs change in response to her judgments of expected utility. And it is suggested by Ismael's appeal to what occurs at the 'conclusion of a piece of deliberative reasoning' (2012, p. 154). If we adopt this interpretation, Tamsin will not have epistemic freedom when she forms beliefs based merely on her partner's prediction. Such beliefs would not count as decisions, and so would not be self-constituting. Tamsin will only have epistemic freedom if she uses the prediction as a guide for what options are best. So she will not generally have epistemic freedom over evidence-based beliefs, and the objection fails.[76] Note that this response is not available to Velleman. Under Velleman's account, introducing

[76] My thanks to an anonymous referee at the *Australasian Journal of Philosophy* for pressing this response.

further conditions on what counts as a decision does not change an agent's epistemic freedom. But under Joyce's and Ismael's accounts, further conditions *can* change an agent's epistemic freedom. It seems Joyce and Ismael can avoid my objection, and deliver an account of epistemic freedom that is sensitive to how beliefs are formed.

If Joyce and Ismael intend the above reading, it is an important clarification of their views. Joyce tells us that the constitutive relationship between decisions and beliefs about them 'ensures that any belief of the form "I decide to do A" adopted during deliberation will be self-fulfilling' (2002, p. 97). But it turns out that not all beliefs formed while deliberating count. Ismael claims that epistemic freedom is unavoidable for any system 'that is representing activity that includes its own' (2007, p. 5), and that the only relevant feature is that the acts 'have reflexive representational content' (2012, p. 156). But if only beliefs formed in certain ways count as decisions, then it's not merely the reflexive content of such acts that gives us epistemic freedom over them. It's how we form them as well.

However, a response that appeals to conditions on belief-formation faces problems of its own. I'll raise two. Firstly, such responses risk making decision too normatively demanding. It seems we can form decisions in irrational ways that aren't responsive to what we take our best options to be. But under the proposed response, we won't be forming decisions at all unless we form them appropriately. This is not a concern for giving idealised models of deliberation—only for using such models to say what decisions actually are. Secondly, the initial objection may re-emerge. Idealised models of deliberation allow for cases where no single option is judged better than alternatives. In such cases, an agent can still decide—perhaps judging the option to be no worse than alternatives. Imagine a case where Tamsin judges that both going to bed and staying up are among her best options. She might then form the belief that she decides to go to bed on the basis of her partner's prediction. If she does this while judging the option is among her best, it will count as her decision. So, again, Tamsin will have epistemic freedom over an evidence-based belief.

Are there other ways that Joyce, Ismael, and Velleman might respond to the initial objection? They could introduce prior ignorance conditions on deliberation. This is another way in which ignorance accounts avoid the problem. In the last chapter, section 2.5, I argued that ignorance conditions should be satisfied by agents not having evidence of how they will decide. If agents do not have evidence of their decisions or actions while deliberating, they will not form beliefs based on such evidence. But unconstraint accounts aim to avoid ignorance conditions. They take agents to form beliefs unconstrained by evidence.[77]

Another way defenders of unconstraint accounts might respond is by claiming that an agent's epistemic freedom comes from her being unconstrained by her evidence. It doesn't matter whether she then follows the evidence or not. But this response is insufficient. To see why, imagine an extreme case. Tom has psychological mechanisms in place that guarantee he *only* forms beliefs based on evidence, and knows this about himself. It may still be true that if Tom were to form a belief in an act or decision he doesn't have evidence for, it would be self-fulfilling or self-constituting. He can't form the belief, but if he did, he'd be justified in it. Tom would count as unconstrained by his evidence and as having the same kind of epistemic freedom we do. Yet we would not expect Tom to experience apparent freedom. We shouldn't have freedom over evidence-based beliefs. And Tom shouldn't have epistemic

[77] Again, one could adopt the hybrid view described above (section 3.2) and begin with ignorance conditions and add in unconstraint. But while such a hybrid view avoids my first objection, it doesn't avoid a second.

freedom at all. The fact that unconstraint accounts get the wrong results here points to an underlying concern with these accounts—one I'll turn to in section 3.5.

### 3.3b Forming Beliefs without Justification

There is a second even more serious concern. If unconstraint types of epistemic freedom are to explain our apparent freedom in deliberation, they must account for some sense in which we seem able to form beliefs unconstrained by evidence. Velleman attempts to do so, by arguing that we are epistemically *licensed* to form beliefs unconstrained by evidence. And both unconstraint accounts might seem to imply we're epistemically justified in *forming* beliefs unconstrained by evidence—not merely justified in the beliefs once they're formed. This justification could explain apparent freedom. But, I'll argue, we're not justified in forming beliefs unconstrained by evidence. So such justification can't explain apparent freedom.

Here's how the explanation is supposed to go. Velleman, Joyce, and Ismael claim that self-fulfilling and self-constituting beliefs are justified once formed. Once Tamsin has formed the belief that she'll (decide to) go to bed, she has evidence of its truth, because she knows the belief is self-fulfilling (or self-constituting). This justification is meant to give her 'license' to form such beliefs now (Velleman 1989b, p. 74 and passim), imply she is not "'hemmed in'" by her evidence (Joyce 2002, p. 98) and make her evidence of what she will do 'irrelevant' (Ismael 2012, p. 154). But these further steps are far from trivial. As Grice (1971, pp. 267–8), Langton (2004), and Setiya (2008, pp. 400–1) all argue, agents are not justified in forming beliefs in the absence of adequate grounds, even if they know their beliefs will be justified once formed. Nor are agents justified in ignoring their evidence. These points have been persuasively argued before, so I will be brief. Justification to form beliefs depends on what epistemic grounds agents have—not on what evidence they would have, were they to believe differently and have different evidence. Before Tamsin decides, she doesn't have grounds for thinking that she'll go to bed *and* grounds for thinking that she'll stay up (and similarly for the corresponding decisions). She is not justified in forming either belief. At most she has grounds for one of these. She may be justified in either belief once she forms it. But that's a different matter. So the justification she has before she decides can't explain her apparent freedom. Ignorance-based explanations of apparent freedom avoid this problem, because they don't appeal to this justification.

Consider the following case, a variant of one discussed by Langton (2004, pp. 257–8). I'll present this case as a counterexample to Velleman's account, and then consider its relevance for Joyce's and Ismael's accounts. Tamsin's son Carlo is thinking about what present he will unwrap at Christmas. He has inductive evidence from previous years that any belief about what he'll unwrap (within reason) will be justified once formed—when he forms the belief that he'll unwrap a toy bear, he does, and so on for other gifts. Moreover, these beliefs are self-fulfilling—if he didn't form the belief that he'll unwrap a toy bear, his mother wouldn't know to buy him one, and he wouldn't unwrap one. Carlo is licensed, in Velleman's sense, to form various incompatible beliefs about what he will do. These beliefs are self-fulfilling and Carlo has evidence they will be justified once formed. And yet it seems Carlo is *not* justified in forming these beliefs. His epistemic grounds do not license him to. So we can't appeal to this justification to explain why agents experience apparent freedom in deliberation.

Should we ascribe the fault to Carlo's mistaken beliefs about Santa? No. Enlightened-Carlo, who has realised the role his mother plays in making his beliefs self-fulfilling still isn't justified in forming them. Does the fault lie in the fact that the relation between Carlo's belief and its truth is causally mediated by Carlo's mother buying the gift? No. The relation between decision and action is always causally mediated—and Velleman's account explicitly builds in causal conditions from the start.

Langton (2004, p. 245, passim) and Setiya (2008, pp. 401–2, passim) describe similar cases as involving 'faith' and 'wishful thinking', where agents go *beyond* their evidence in forming beliefs. But it's worse. Under Velleman's account, agents are not only licensed to go beyond their evidence, but to contradict and go against their evidence. Even if an agent has complete evidence that entirely settles what he will do, he is still justified in believing otherwise. Even if 'the agent is confronted by all of the evidence about his future action, and…fully appreciates its evidentiary force…This complete evidence cannot require the agent to draw a particular conclusion about what he'll do' (1989a, p. 153). Nor is it only external evidence that agents are unconstrained by. Agents are unconstrained by even complete knowledge of their own motivations. Evidence of motives that 'constitutes compelling evidence for expecting (an agent) to perform one action rather than another…still cannot compel the agent himself' (ibid., p. 150). Given how strong unconstraint needs to be, there is even less reason for agents to be justified in forming beliefs unconstrained by evidence.

Does Carlo's case also create problems for Joyce's and Ismael's accounts? It does. Initially it might not seem so. After all, Carlo's beliefs are causally self-fulfilling, not self-constituting, and the latter is what Joyce and Ismael appeal to. But things are not so simple. While I have been careful to distinguish the two types of unconstraint account, both Joyce and Ismael take agents to have epistemic freedom regarding both *causally* self-fulfilling and self-constituting beliefs. Even though they only take self-constituting beliefs to be relevant to our epistemic freedom over decisions, they employ a generic notion of 'self-fulfilling belief' that includes both. These are beliefs that 'if the agent has them then they are true' (Joyce 2002, p. 95), or beliefs whose truth is 'probabilistically dependent' on the act of forming the belief (Ismael 2007, p. 6). The same kind of unconstraint that applies to self-constituting beliefs is meant to apply to causally self-fulfilling beliefs. Nor can I see a principled reason to restrict unconstraint to self-constituting beliefs. Carlo's case is enough to trouble Joyce's and Ismael's accounts.

It may seem that Joyce and Ismael avoid the worry in another way. Perhaps it doesn't matter whether agents are justified in forming beliefs unconstrained by evidence. Perhaps it is enough that the beliefs are justified once formed, or that agents don't do anything *wrong* in forming them. But if epistemic freedom is to explain why certain options appear available to be decided on, we need more. We need to identify some positive sense in which we seem able to form the beliefs. Velleman's appeal to epistemic license seemed to provide this. If Joyce and Ismael give up this appeal, they give up Velleman's purported explanation of apparent freedom. As it stands, if agents are not justified in forming decision-beliefs, we have no account of how agents appear *free* to form them.

Another way defenders might respond is by arguing that our ordinary epistemic intuitions are out of place when it comes to self-fulfilling or self-constituting beliefs. Perhaps we are justified in forming these beliefs unconstrained by evidence, but forget that we are. Or perhaps we should revise our notion of justification to accommodate these beliefs. In response, note that the intuitions I'm appealing to aren't weighty. I'm not assuming we *infer* beliefs based on *evidence*, but only that our justification in forming beliefs is given by our general epistemic grounds, whether or not we actually infer. Nor is it a matter of neglect. Philosophers have considered these unusual cases and remained unconvinced. There are ways to convince sceptics or successfully advocate revision of our concepts. One can appeal to structurally similar cases outside the disputed area. But Carlo's case raises problems for such appeals—structurally similar cases fall the other way. One can also appeal to theoretical advantage—perhaps we should accept the relevant notion of justification because it explains apparent freedom. But, as I will argue, there is no such advantage—once we've introduced what's needed for unconstraint accounts of epistemic freedom, there are easier ways to explain apparent freedom.

### 3.4 Underlying Trouble

Unconstraint accounts imply that we experience apparent freedom even over evidence-based beliefs. And they rely on the mistaken idea that we can justifiably form beliefs unconstrained by evidence. In this final section, I argue for a diagnosis of what has gone wrong—unconstraint accounts presuppose an unexplained apparent ability to form different beliefs. Only with this apparent ability do these accounts give us an explanation of apparent freedom. But this is an apparent ability that makes epistemic freedom redundant.

I've argued that agents aren't justified in *forming* beliefs unconstrained by evidence. But one might think that an agent's apparent freedom is explained by the justification an agent *would have*, were she to form the belief. This would avoid my second objection. However, to appeal to this conditional justification, we'd need to identify a sense in which the agent takes it she *could* or *might* form such beliefs—minimally, a modality in which such beliefs appear possible.

Only with this additional modality might the justification she has once these beliefs are formed explain her apparent freedom now.

To see why some kind of modality is required, consider an agent who knows a belief would be self-fulfilling, were he to form it, but who correctly takes it he *cannot* form the belief. Unconstraint accounts need to rule out such cases. Imagine Raamy, who knows that if he were to believe he will swim into a river teeming with crocodiles, he will indeed swim into the river, given how stubborn he is about acting as he believes. He is Velleman's agent *par excellence*, strongly motivated to act as he expects. But Raamy also knows he is unable to form the belief. He has such an overwhelming fear of crocodiles that he can't even *imagine* swimming into the river, let alone believe he will. The mere thought of crocodiles fills him with dread, so that any attempt to conceptualise, visualise, or form a belief about swimming into the river fails. Raamy won't feel free to decide to swim into the river. But if it were merely the justification he would have were he to form the belief that explained his apparent freedom, he would feel free.

Or consider Sunny. Sunny knows she lacks the concept eigenvalue. So she knows she is unable to form the belief that she will compute the eigenvalues for a simple matrix in front of her. When her teacher asks her to calculate the eigenvalues, she doesn't know what she's being asked. But she is familiar with matrices and adept at arithmetic. If she were to form the belief that she will compute the eigenvalues, this would imply her having the concept eigenvalue. Then she would indeed be able to compute the eigenvalues, and, being a diligent student, would do so. She'd be justified in the belief if she formed it. But in her present state, she would not appear free to decide. Like Raamy and Tom (who can only form beliefs based on evidence), the mere justification she would have, were she to form the belief, would incorrectly attribute her apparent freedom now.

I suspect unconstraint accounts of apparent freedom rule out such cases by implicitly relying on an additional modality in which agents are or appear free to form beliefs. What sort of modality might they appeal to? Could they appeal to an apparent 'metaphysical' freedom to believe in ways not causally or nomologically determined by prior states of affairs? Not likely. This would make apparent freedom mysterious, and take away whatever advantage came from appealing to ordinary beliefs. They could try an epistemic modality—the fact that various beliefs are compatible (or thought to be compatible) with the agent's other beliefs. But such an approach places ignorance conditions on deliberation, which unconstraint accounts aim to avoid.

Sometimes Velleman seems to appeal to the fact that some beliefs are 'wishful' and desire-responsive to explain why they appear under an agent's control (1989a, pp. 69, 127). Sometimes he notes the fact that agents don't form them in response to evidence (ibid., p. 25). And sometimes these ideas are combined: 'we believe what we believe because we want to, not because the evidence would have dictated it' (ibid., p. 162). But these appeals won't explain why agents appear free to form the beliefs. Regarding evidence, precisely what is in question is how an agent takes it she *can* form beliefs unconstrained by evidence. We shouldn't use an unexplained ability to do so to explain this possibility. On the desire side, things fare no better. Velleman claims agents can appear free to form beliefs that are *not* in keeping with their greatest desires, or what they take them to be (ibid., pp. 145, 150). So apparent freedom can't be explained by desire-responsiveness.

Could Velleman revise his claim, and require the relevant beliefs to be desire-responsive? Perhaps. He could also adopt a more sophisticated condition, and require beliefs to be responsive to what an agent takes the expected utility of her acts to be—a possibility suggested by Joyce's account. Perhaps this is what explains why she seems free to form various beliefs. But the problem is that if desire-responsiveness explains why she appears free to form these decision-beliefs, why bother with the detour via epistemic freedom? What does it matter if she's justified in her decision-beliefs once she forms them? The epistemic part of the account no longer does any work in explaining apparent freedom (even if it serves other roles). We may as well explain apparent freedom by the fact that we form decisions in response to desires or judgments of expected utility.

Finally, we might try appealing to a real or apparent primitive agent-ability. Sometimes Velleman claims that self-fulfilling beliefs are formed 'spontaneously' (1989a, p. 25), 'voluntarily', or 'at will' (ibid., p. 66). But appealing to such primitives is not a promising approach. Beliefs are usually taken to be involuntary states that we simply find ourselves in. While it may turn out that we can form beliefs at will, we shouldn't appeal to an unexplained ability to do so to explain apparent freedom. The more reasonable way of interpreting Velleman's account is that the appearance of forming beliefs at will is to be *explained* by epistemic freedom—for support, see Roessler (2013, p. 43).

Ismael speaks of the decision process as something 'I *implement*', which overrides previous evidence (2012, pp. 154–5). And she sometimes speaks of the 'representational acts' and 'activity' of the deliberator (2007). But these descriptions give us no insight into why the decision process counts as something 'I' implement, or why forming beliefs appears as an activity at all—and so why an agent might appear free to form different beliefs. Even if we think of deliberation as merely a changing stock of evidence, we still haven't identified why such changes appear possible.

I suspect Velleman, Joyce, and Ismael are relying on some combination of desire-responsiveness and a primitive apparent psychological ability to form different beliefs. If agents appear able to form different beliefs (unconstrained by evidence), and yet still end up with justified beliefs, this might explain their apparent freedom. But even if we grant a primitive apparent ability, epistemic justification no longer plays a role in explaining apparent freedom. Epistemic freedom becomes redundant. Once agents have a primitive apparent ability to form different beliefs, this is just a primitive apparent ability to decide in different ways. No further explanation of our freedom in deliberation is required. Justification may still play a role in explaining why these states count as beliefs. But to explain apparent freedom, we have no reason to adopt an unconstraint account over an account that simply grants us a primitive apparent ability to decide. Ignorance accounts avoid this problem because they appeal to epistemic possibility and other epistemic conditions to explain how an agent seems free to decide in different ways.

### 3.5 Fichtean Insights

If I'm right, identifying decisions as beliefs does not lead to a plausible unconstraint account of our apparent freedom in deliberation. Even if agents can form self-fulfilling or self-constituting beliefs unconstrained by evidence and do nothing *wrong* in forming them, epistemic features do not explain how agents seem *free* to form them. If identifying decisions as beliefs does not explain our apparent freedom in deliberation, we might expect to find other accounts that also identify decisions as beliefs, but that don't take epistemic features to explain apparent freedom. There are such accounts. I end this chapter by briefly considering the lessons we can draw from them.

Some accounts that identify decisions as beliefs are found in contemporary philosophy of action. Hampshire and Hart (1958) take decisions to entail beliefs that are *not* formed on the basis of evidence, but on the basis of (practical) reasons. But, unlike defenders of unconstraint accounts, they don't take features of these beliefs to *explain* our apparent freedom to decide. Pears (1968) comes close to taking intentions and decisions to be partly constituted by beliefs. But he doesn't think we form these beliefs unconstrained by evidence, but based on knowledge of our desires. And he doesn't take our justification to form these beliefs to explain our apparent freedom. Pears, Hampshire and Hart primarily use relations between decision and belief to explain our knowledge of our future actions. On none of their accounts do epistemic conditions explain our apparent freedom to decide.

Another example of an account that identifies decisions as beliefs comes from a rather surprising place—the German Idealist tradition. Johann Gottlieb Fichte argues for an equivalence between thinking and willing in choice. His account is a particularly neat illustration of why identifying decisions as beliefs does not explain our apparent freedom in deliberation. Fichte begun his philosophical career as an interpreter of Kant, agreeing with much of Kant's system. But he believed Kant's first two critique had wrongly separated theoretical reason (reason about what is the case) from practical reason (for Fichte, reasoning about what *ought* to be the case)—suggesting a disunified picture of reason.[78] Fichte attempted to provide a unified account of practical and theoretical reason, particularly in his *Wissenschaftslehre nova methodo* ([1796/99] 1992, hereafter Wnm).[79]

There are various ways in which Fichte attempted to unify practical and theoretical reason. The most important for our purposes is that Fichte aimed to show that practical activities were equally (in some sense) theoretical activities, just looked at from a different point of view. Along the way, he argued that this unity between practical and theoretical reason was actually demonstrated in consciousness, by an equivalence in deliberation between acts of willing and acts of thinking of willing. Effectively, Fichte argues for an equivalence between decisions and beliefs in decisions, anticipating Joyce's and Ismael's claim that decisions are self-constituting beliefs. However, unlike Joyce, Ismael and Velleman, Fichte does not take the equivalence between decision and belief to support a reduction of decision to belief, or an explanation of freedom in epistemic terms. In fact, Fichte couldn't be clearer about the kind of radical freedom that his account presupposed.

Here's how Fichte conceives of deliberation. Deliberation involves a kind of openness and uncertainty, as the 'I' (roughly the rational mind, or reason itself) oscillates between various options. When decision occurs, the oscillation ceases, and the I becomes fixed on a single option. 'In an act of willing, I seize control of my wandering thoughts and restrict them to a single point' (Wnm, p. 126). This is 'a process of voluntary limitation, inasmuch as one

[78] Fichte would also be critical of the perspectivalist accounts (chapter 2, section 2.5) that separated practical and theoretical perspectives.

[79] The title roughly translates as 'Theory of Scientific Knowledge after a New Method', 'science' indicating all systematic forms of knowledge. Fichte presented this account in lectures, repeated over the winter semesters 1796–1799, two transcripts of which have survived: Kraus and Halle (see editor's introduction). Page references are to the Kraus transcript, except where noted.

focuses one's will upon a single new object' (Wnm, p. 124). Crucially, this restriction is equally an act of thinking and an act of willing: 'Through the act of thinking of willing, I will; and because I will, I think of the act of willing' (Wnm, Halle p. 114). Acts of thinking and willing appear as equivalent: 'the act of willing is not separated within consciousness from the act of thinking of willing' (Wnm Halle, p. 15). As philosophers observing the activities of the I we can separate the two acts, but for the I they appear as one (Wnm, pp. 182–3).

Much more would need to be said in order to lay out the full workings of Fichte's system. I use this brief outline to make only two points. Firstly, taking decisions and beliefs to be equivalent is not enough to reduce one to the other. Velleman explicitly uses an equivalence of belief and decision to reduce practical to theoretical reason, part of his general project of explaining deliberation in epistemic terms (1989a, pp. 10–11). But unless one has prior reason to be suspicious of decisions, even a complete functional equivalence between them is not enough to imply that one reduces to the other. Fichte uses an equivalence between decision and belief to support a unity of practical and theoretical reason, not reduction.

Secondly, and more importantly for our purposes, an equivalence between decision and belief does not provide an unconstraint explanation of our (apparent) freedom in deliberation. We still need to account for a freedom to believe. Fichte for the most part agrees with Velleman, Joyce and Ismael that decisions are beliefs (given how things appear in consciousness). And he agrees that we do nothing epistemically *wrong* in forming such beliefs—he is concerned throughout with distinctly *rational* activities of the mind. But Fichte also explicitly presupposes that the I is free to act in various ways. This freedom is at work in all rational acts of thinking, believing and willing: 'Your thinking is, for you, an *acting*'

([1797/98] 1994, p. 522). Fichte indeed characterises the I as a kind of spontaneous rational activity, an activity not governed by mechanistic laws ([1797/98] 1994, p. 439).[80] And it is precisely this freedom that is exercised when the I settles on a single option in deliberation. The movement from oscillation to settling may be thought of in practical terms (as the willing of an action) or theoretical terms (as thinking of a willing). But in either case, the I must be free to act in the relevant way.

I'm not arguing that Velleman, Joyce and Ismael presuppose that agents are radically free in the way Fichte supposes them to be. But they must account for some sense in which agents are, or appear, free to form different beliefs. Fichte's account highlights the explanatory gap in unconstraint accounts. And the need for a further kind of freedom undercuts the attempt to explain apparent freedom merely in terms of our *justification* to form different beliefs. If we grant agents a freedom to form different beliefs, and beliefs are decisions, no further explanation is required to explain our apparent freedom to decide in different ways.

That said, there are ways in which Velleman, Joyce and Ismael do appeal to similar kinds of freedom as Fichte. While Velleman claims agents *don't* have metaphysical freedom (1989b, p. 73), as we saw, he thinks agents will seem free to form beliefs that go against complete evidence that entirely settles what they will do (1989a, pp. 150, 153). This is, at least, very close to an *apparent* metaphysical freedom. I also noted above (section 3.4) that Joyce and Velleman might appeal to the fact that agents form beliefs in response to their desires, or what they judge their best options to be, in order to explain their apparent freedom to

[80] These quotes are from the published introductions Fichte revised out of his 1796/99 lectures. Similar remarks are found in the lectures, where Fichte contrasts his own idealism with 'dogmatism', the attempt to explain rational activities of the mind using mechanistic laws (Wnm Halle, pp. 21–2).

believe. Fichte's account suggests an additional reason for rejecting this route. Fichte uses our responsiveness to practical reasons to unify practical and theoretical reason in something *practical*. Responsiveness to judgments about practical norms (or desires) is, if anything, characteristic of practical reason. It doesn't explain our practical freedom in theoretical or epistemic terms. Overall, while Fichte's account shares important features in common with unconstraint accounts, it helps demonstrate why unconstraint accounts don't explain apparent freedom in deliberation in epistemic or theoretical terms.

## Conclusion

In chapter 2, I presented an epistemic model of deliberation based in ignorance conditions. An agent must be ignorant of what she will decide and do if she is to reasonably deliberate. And I argued that the model was enough to characterise deliberation and explain an agent's apparent freedom in deliberation. This model faces attacks from various quarters. Competing ignorance accounts argue that agents can properly have no degrees of beliefs in what they will decide and do as they deliberate (Levi, Price). I've argued against this requirement. Even if such degrees of belief are hard to measure, they serve a useful function, and agents should not forgo them. Competing perspectival accounts argue that agents presume their decisions have no causal history, or are otherwise evidentially severed from states of affairs causally upstream (Price). But these accounts rely too closely on causal notions to be suitable for explaining causation and causal asymmetry.

The most serious challenges to my ignorance-based model of deliberation come from unconstraint accounts (Velleman, Joyce, Ismael). According to these accounts, agents are justified in forming different beliefs in what they will do or decide unconstrained by evidence. This is what explains their apparent freedom. These accounts are competitors to ignorance accounts (*contra* Price). However, they turn out not to explain our apparent freedom in deliberation in epistemic terms. They mistakenly imply agents have apparent freedom even over evidence-based beliefs. And they rely on a mistaken notion of justification—agents aren't justified in forming beliefs unconstrained by evidence. Underlying these troubles, unconstraint accounts presuppose an apparent ability to believe in different ways. But, under these accounts, this just *is* an apparent ability to decide in different ways. If so, epistemic unconstraint does no work in explaining our freedom to decide. Altogether, if we're to explain causation by appeal to deliberation, the epistemic model remains the most promising epistemic account. In the following two chapters, I turn to the project of building the epistemic model of deliberation into a deliberative account of causation.

# Chapter 4

## Accounting for Causation

*…man looks at nature from his own point of view; not the point of view of a thinker, anxious to find out the truth about nature as it is in itself, but the point of view of a practical agent, anxious to find out how he can manipulate nature for the achieving of his own ends.*
R. G. Collingwood (1940, p. 310)

In chapter 1, I introduced Russell's challenge: given that causes do not feature in fundamental physics, and do not necessitate their effects, how should we understand causation? In this chapter, I answer this challenge. I use the epistemic model of deliberation developed in chapter 2 to construct a deliberative account of causation, according to which causal relations correspond to the evidential relations we use in deliberation when we decide on one thing in order to achieve another. I defend this account in the form of a biconditional: *A* is a cause of *B* *if and only if* an agent's deciding on *A* for the sake of *B* in proper deliberation evidentially settles *B*.

The chapter proceeds as follows. In section 1, I set out three desiderata on accounts of causation: an account must explain why causation should matter to agents, how agents can apparently intervene to make use of causal relations, and how causation relates to fundamental laws. In sections 2 and 3 I argue that prominent interventionist and reductive counterfactual accounts of causation do not adequately satisfy these desiderata. In section 4 I develop the deliberative account of causation and argue that it satisfies these desiderata. Causation is useful because causal relations correspond to the evidential relations required for effective decision-making. Agents justifiably take it they can intervene because their evidence leaves open how they decide, yet their decisions evidentially settle their results. And causation relates to fundamental laws because fundamental laws are required for evidential reasoning. In section 5, I go on to consider the metaphysical upshots of the deliberative account. I argue that the account need not reduce causal relations to evidential relations. But it does constrain first-order metaphysical accounts and can justify why the relations they pick out deserve to be called causal. The deliberative account can in fact be used to support a variety of projects, while still providing, in itself, a philosophically rich account of causation. In section 6, I argue that the deliberative account also avoids problems with circularity that have troubled other agent-based accounts.

### 4.1 Three Desiderata on Accounts of Causation

What do we need from an account of causation? In chapter 1 I considered a challenge from Russell (1912–13): when we look at how fundamental physical theories are formulated, causal terms don't feature. What we have instead are dynamical equations relating whole states of systems across times—causal information is neither required by the theory nor directly identifiable from the theory. I took the important upshot of Russell's challenge to be that while we might not want to eliminate causal talk altogether, we need to look somewhere other than predictions and derivations of fundamental physics to understand how causation fits into a scientific picture of the world.

One influential response to Russell's challenge has been to tie causation to counterfactual reasoning, effective manipulation and control. Causes are *responsible* for their effects, in the sense that if the cause hadn't occurred, the effect would not have occurred either, and so we can manipulate an effect by manipulating its causes. While not all causes need be manipulable in this way, causal talk has its natural home in our interactions and interventions in the world: not in observing things happen, but in making them happen.

This line of thought is found in early manipulationist accounts of causation, including those by R. G. Collingwood, Georg von Wright and Douglas Gasking. I began this chapter with a quote from Collingwood, in which he directly ties one sense of causation to taking up 'the point of view of a practical agent, anxious to find out how he can manipulate nature for the achieving of his own ends' (1940, p. 310). Von Wright develops this idea into a formula relating causation and manipulation: '*p* is a cause of relative to *q*, and *q* an effect relative to *p*, if and only if by *doing p* we could bring about *q* or by suppressing *p* we could remove *q* or prevent it from happening' (1971, p. 70, my emphasis). Gasking (1955) defends a similar view.

More recent defenders of manipulationist accounts make less use of agential notions, like action and ends, but still maintain the importance of causation for making practical decisions. Cartwright, for example, argues that causal notions are ineliminable because 'they are needed to ground the distinction between effective strategies and ineffective ones' (1979, p. 420). Similarly Woodward, whose account I turn to shortly, claims 'our interest in causal relationships and explanation initially grows out of a highly practical interest human beings have in manipulation and control' (2003, p. 10). Understanding causation in terms of manipulation, control and counterfactuals has presented itself as a promising route to answering Russell's challenge.

Moreover, it seems that understanding the practical import of causation is not just one way to answer Russell's challenge. Given the way we use causal information in practical and scientific life, an account of causation owes us an explanation of how causation plays this role. We use causal information in order to decide polices, behaviour and strategies—we work out what kind of effects our actions will have in order to decide on the right actions. We need to distinguish causal relationships from those that merely allow us to make good predictions. This is why we have the familiar refrain "correlation does not imply causation". An account of causation needs to address Russell's challenge in a way that makes sense of the practical import of causation. Hartry Field even describes answering this challenge as 'probably the central problem in the metaphysics of causation' (2003, p. 443).

With these points in mind, here is a first desideratum on accounts of causation: an account should make sense of how causation and knowledge of causal relations help agents achieve their ends. I call this the 'pragmatic requirement'. While causation might play other roles, it must be relevant for agents attempting to manipulate the world. This pragmatic requirement is similar to Mellor's means–end aspect of causation (1988, p. 231) and shares features in common with Price and Weslake's 'practical relevance constraint' (2009). Note that the pragmatic requirement doesn't imply we can make use of every causal relation. It requires causation and knowledge of causal structure to be useful in general.

This first desideratum leads natural to a second: an account of causation should explain how causation is *accessible* to agents—how agents can (or justifiably take it they can) intervene in the world so as to manipulate causes. Causation should not be a useful but inaccessible tool. Whatever notion of intervention is appealed to in an account of causation should be relevant for what agents can do. A counterfactual account, for example, must explain how agents can (or justifiably take it they can) bring about counterfactual antecedents. The desideratum is partly motivated by concerns about agency raised in chapter 1 (section 1.2). Given that fundamental laws perfectly account for the movements of everything at the fundamental level, how does it seem we can intervene in the world? If it were obvious how agents could intervene on causes, we wouldn't need an explanation. But it's not obvious at all.

A third desideratum arises directly from Russell's challenge: accounts of causation should explain how causation relates to fundamental physical laws. Russell argued that causation has no clear place in fundamental physics. So it seems we have two different systems for describing the world: one in terms of physical laws, the other in terms of causes. Given these systems are about the very same world, we should be able to say something clear and perspicuous about how they relate. At the very least, we should be able to show that law-based and cause-based systems are compatible, such that we won't be lead astray by using both kinds. This desideratum could be met by explaining causation in terms of laws, or vice versa—we'll see examples of both kinds of strategy below.

Overall, an account of causation should meet three desiderata. It should explain:

1) how causation is useful to agents (the pragmatic requirement);
2) how causation is accessible to agents;
3) how causation relates to fundamental laws.

### 4.2 Interventionist Accounts

I'll now consider how some prominent accounts of causation fare with respect to the three desiderata. I begin with non-reductive interventionist accounts. These accounts make sense of causation in terms of manipulation and control—causes are a means of manipulating effects. These accounts are also explicitly non-reductive—they don't aim to metaphysically or conceptually reduce causal relations to other relations. Defenders of this approach are at best agnostic on whether such reductions can be achieved, and argue causal concepts require neither explication nor grounding in non-causal relations. The focus is instead on exploring the connections between causal relations, counterfactuals and interventions.

### 4.2a Pearl's and Woodward's Interventionist Accounts

I'll consider accounts from Judea Pearl (2000) and James Woodward (2003) and take them to be paradigmatic of interventionist accounts.[81] We begin by representing the causal relata of a system using variables. The values these variables take could represent the occurrence or obtaining (or not) of *particular* events or states of affairs (Suzy's actual throwing of the rock here and now). This may allow one to derive claims about singular or token-causation (see, for example, Woodward 2003, pp. 76‒86). But more commonly the values represent generic states (Suzy's throwing of a rock), which can only allow us to derive type-causation. We then make some causal assumptions about how different variables in the system relate to one another. For example, we might begin by assuming that A (a variable that takes the values 0 or 1, depending on whether Suzy throws a rock) is a direct cause of B (a variable that takes values 0 or 1, depending on whether a nearby bottle smashes). We then represent this causal relation using a 'structural equation', such as:

[81] Cartwright (1979; 2003) and Spirtes, Glymore and Scheines (1993) also defend interventionist accounts, with slightly different technical details and metaphysical underpinnings.

$$B := A$$

This equation tells us that if A were to take the value 0 (Suzy doesn't throw a rock), B would also take the value 0 (the bottle doesn't smash). If A were to take the value 1 (Suzy throws a rock), B would also take the value 1 (the bottle smashes). But the equation does not represent simply an equality between the values. The equation is asymmetric—this is why I have used the notation ":=". A *causally determines* B, rather than the reverse. These structural equations capture qualitative or quantitative information about how B would change as a result of changes in A—although I will simplify to binary variables throughout. The qualitative information of the equation can also be represented using directed acyclic graphs. Figure 1 represents A as a direct cause of B:

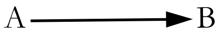


Figure 1: Causal graph showing A as a direct cause of B.

So far all we have is a single direct causal relation. The power of the interventionist approach comes when we consider more complex causal structures. Say Kyle has observed a correlation between C (his celebrating a birthday), D (his drinking alcohol) and E (his feeling euphoric). But he's not sure whether this is due to direct causal relations between C and D and D and E (so that drinking causes his euphoria), or whether C is a common cause of both D and E (so that celebrating causes both drinking and euphoria). That is, he is not sure which of the following graphs (figures 2 and 3) represents the causal structure of the system.

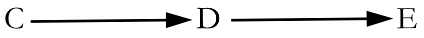


Figure 2: Causal graph showing C as a direct cause of D and D as a direct cause of E.

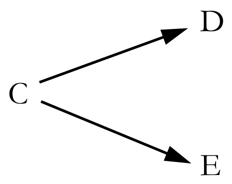


Figure 3: Causal graph showing C as a common direct cause of D and E.

In real life, we often use experiments to determine which of the figures correctly represents the causal structure of a system. Interventionist accounts take the causal structure of experimental testing to capture what it is for a direct causal relation to hold. To show how this works we need to introduce an intervention variable, represented by I. This is a variable (relative to D and E) that directly determines the value of D, through some exogenous causal process—a process that was not part of the original causal system. For example, Kyle might toss a coin toss to determine whether he drinks or not. This intervention 'breaks' the causal relation that held between C and D, so that it is now I (the coin toss) that determines the value of D, rather than C (his celebrating).[82] The intervention is also 'surgical', in the sense that it leaves other causal relations intact.[83] Figures 4 and 5 show how interventions are introduced on D, given the causal structures represented in figures 2 and 3.

[82] One can also introduce functional or probabilistic interventions, where the value of I only partially determines D, or determines its probability distribution—see Pearl (2000, p. 70 ff.).

[83] Pearl and Woodward have slightly different requirements on interventions. Say we introduce an intervention on X (relative to Y) to see whether X causes Y. Woodward allows other causal relations to be changed, but requires the intervention to be *statistically* independent of any variable Z that causes the Y and that is on a causal path that does not go through X (2003, p. 98). Pearl requires the intervention variable on X (with respect to Y) to be *causally independent* of variables that cause Y (that are not on a directed path between X and Y) (2000, p. 70). Woodward's complaint is that this condition takes qualitative causal relations between X and Y for granted (2003, pp. 38, 110).

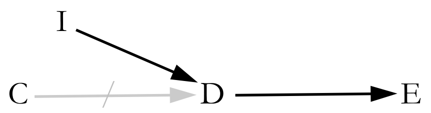


Figure 4: Causal graph showing an intervention on D, given the causal structure in figure 2. The causal relation between C and D is now broken.

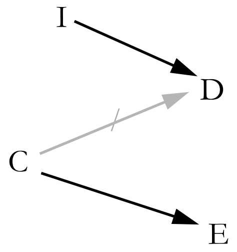


Figure 5: Causal graph showing an intervention on D, given the causal structure in figure 3.

Say Kyle wants to know whether he can experience euphoria more often by celebrating more birthdays or by drinking more often. He can use the coin-toss intervention to find out. If his feeling euphoric remains correlated with drinking, even when his drinking is determined by the coin toss, this suggests the causal structure in figure 2. He can drink to feel euphoric. But if his feeling euphoric is no longer correlated with drinking when he intervenes on drinking, this suggests the causal structure in figure 3. He can only celebrate to feel euphoric.

Woodward and Pearl take experiments like Kyle's to be ways of testing the truth of interventionist counterfactuals—counterfactuals about what would happen to one variable (E) if another variable were to be changed by an intervention (D). But it is the interventionist counterfactuals themselves that give us necessary and sufficient conditions on causal relations, and provide elucidation of what it is for D to cause E. While the actual experiments may be evidence for these counterfactuals, we may always get things wrong—particularly since causal assumptions are required to determine what counts as an intervention. Furthermore, interventionist counterfactuals may still be true, and we can still have evidence concerning them, even when the appropriate experiments cannot be done. So there may be causal relations even when interventions have not or cannot be performed. Nor do interventions need to performed by human agents—natural processes can also be interventions, provided they satisfy the appropriate causal conditions.

These interventionist counterfactuals are 'cause-involving counterfactuals' because their meaning and evaluation is causally-laden throughout. In evaluating these counterfactuals, we must introduce intervention processes that are causally characterised, and keep other relevant causal relations intact. The causal aspects of the counterfactuals cannot be eliminated. It is for this reason that interventionist accounts are non-reductive: they do not attempt to reduce causal relations to non-causal ones or elucidate causal concepts in non-causal terms. Nevertheless, these accounts do aim to give non-trivial necessary and sufficient conditions on what it is for any particular causal relation to hold, and, in Woodward's case, conditions that only appeal to causal information concerning other variables.

Woodward and Pearl agree for the most part on the formal aspects of their accounts. But they disagree on the underlying metaphysics. Pearl takes the structural equations to represent autonomous physical causal mechanisms. These are the primitives of his account, not the counterfactuals (2000, pp. 22, 37). It is partly because he commits to this underlying metaphysics that Pearl characterises intervention slightly differently from Woodward.[84] Woodward does not specify any primitives—not interventions, direct causal relations or counterfactuals. Woodward is at best agnostic on the underlying metaphysics. While he

[84] See footnote 82, p. 125.

argues that interventionist counterfactuals are objective, and do not depend on our abilities or psychology, he claims his theory is 'modest and…commits us to no particular metaphysical picture of the "truth makers" for causal claims' (2003, pp. 121–2).

### 4.2b Evaluating Interventionist Accounts

Various concerns have been raised with interventionist accounts.[85] I will only consider how they fare with respect to the three desiderata above. Regarding the pragmatic requirement, Woodward is sensitive to the fact that an account of causation should explain why causation matters to agents (2003, p. 28):

> What is the point of our having a notion of causation (as opposed to, say, a notion of correlation) at all? What role or function does this concept play in our lives? An important part of the appeal of a manipulability account of causation is that it provides a more straightforward and plausible answer to this question than its competitors.

Woodward takes the benefit of causal knowledge to 'have to do with manipulation and control' (2003, pp. 18–9). If causal relations are 'potentially exploitable for purposes of manipulation, there is no mystery about why we should care about them' (2003, p. 7, cf. p. 138). Pearl makes similar claims (2000, p. 337).

Woodward and Pearl claim that interventionist counterfactuals are relevant for manipulation and control. This might explain why we should care about interventionist-causation. But if this explanation is to be anything but trivial, they must give us an *independent* specification of manipulation and control, and show how causation, explicated in interventionist terms, latches on to these. And this is not how the explanation proceeds. Instead, what we have is a

[85] See, for example, Hall (2007), Strevens (2007), Weslake (2006) and Beebee, Hitchcock and Price (2015).

specification of manipulation and control in terms of interventionist counterfactuals, and *the very same* explication of causation in terms of interventionist counterfactuals. Woodward and Pearl simply assume that their particular explications of manipulation and control pick up on useful relations. But if one is worried that interventionist counterfactuals are not latching onto useful structure, it is no comfort to be told that the structure they latch onto is useful *because* it consists of interventionist counterfactuals. The usefulness of these interventionist counterfactuals is precisely what's in question. If manipulation and control are explicated in exactly the same way that causation is, we don't have a sufficient response to the pragmatic requirement. The concern is not that the account involves causes in the counterfactuals, and so doesn't reduce causation to non-causal relations. The concern is that a justification that appeals to the very same causally-characterised types of manipulation and control each time is too small a circle to be philosophically illuminating. It gives us no independent grasp of why causation, understood in these particular causal terms, is useful to agents. Similar points are made in Weslake (2006) and Price and Weslake (2009).

A similar concern arises regarding the second desideratum: explaining how agents can, or justifiably take it they can, intervene in the relevant sense. Interventionist accounts simply take it for granted that agents have (or justifiably take themselves to have) causal powers to intervene in the world, in the sense required to make use of interventionist counterfactuals. Woodward, for example, claims (2003, p. 150):

> Human beings often are (and are justified in believing that they are) in situations in which they can perform actions affecting some variable X meeting the conditions …on interventions set out above, and in which their interest is in knowing what will happen to some other variable Y under such an intervention.

Pearl also presumes agents can intervene when he explains how children acquire causal knowledge: ‘a child can infer that shaking a toy can produce a rattling sound because it is the child’s hand, governed solely by the child’s volition, that brings about the shaking of the toy’ (2000, p. 253). But he doesn’t consider why the shaking of a toy or the child’s volition should satisfy conditions on intervention. Recall, the conditions on intervention put forward by Pearl and Woodward are precise and technical, designed to yield causal relations from statistical data and causal assumptions. Why should we think our ordinary, everyday actions, in pursuit of goals, are likely to satisfy these causal and statistical assumptions? Neither Woodward nor Pearl explains this.

The closest Pearl comes to accounting for the relation between interventions and actions is to simply take agents to have a primitive ability to intervene in the world (2000, pp. 108–9). Pearl distinguishes between *acts*, which are conceived of ‘from the outside’ as mere events in the world with causal antecedents, and *actions*, which are conceived of ‘from the inside’ as options contemplated in deliberation. He claims that *actions* ‘can neither be predicted nor provide evidence since (by definition) they are pending deliberation and tum into acts once executed’ (2000, p. 108). Actions are ‘objects of free choice’ and ‘change the probabilities that acts normally obey’ (2000, p. 109).[86] If agents do have a primitive ability to intervene, this is why ‘actions’ have the unusual evidential relations they do. But he doesn’t explain why actions should be characterised in this way, or how agents have (or take themselves to have) primitive abilities to intervene. So his account doesn’t satisfy the second desideratum.

Perhaps the strongest response Woodward and Pearl can make to the pragmatic requirement and the second desideratum is to appeal to experimentation. It is, after all, through

[86] On this point, Pearl comes very close to Price’s perspectivalism—see chapter 3, section 3.1d.

experiments that we come to have evidence of interventionist counterfactuals. As far as we take experiments to involve instances of manipulation and control, we have reason to think that the counterfactuals we reason to using experiments also pick up on relations of manipulation and control. And so we must be able to intervene in the way required to test for causal claims. Here I am reversing the order of explanation in Woodward's account (2003, p. 35). But this is at most a transcendental justification that the structure is to some degree useful and accessible. It gives us no direct insight into why this particular counterfactual structure is relevant.

Furthermore, without a deeper understanding of what is going on in experiments, it is unclear why the causal claims extend from experimental situations in which agents do intervene, to situations in which they don't or can't. What reason do we have to think that experimentation latches onto a structure that holds throughout? While Woodward, for example, acknowledges that the interventionist account rests on there being an appropriate domain in which we can evaluate interventionist counterfactuals (2003, p. 314), he offers no explicit guidelines what constitutes such a domain. It might seem that Pearl answers this challenge because he is clearer about the metaphysical underpinnings of his theory. Perhaps the existence of fundamental autonomous causal mechanisms can explain how experimental testing leads to generally useful causal relations. Pearl claims that we expect causal relations to be stable because they are 'ontological, describing objective physical constraints in our world' (2000, p. 25). But, once again, since these fundamental causal mechanisms are specified entirely in terms of interventionist counterfactuals, we have no independent justification for their use. So we still don't answer the pragmatic requirement.

To answer my concerns, it would help interventionist accounts if they could secure a connection between interventionist counterfactuals and fundamental laws. After all, fundamental laws are typically taken to hold universally and support counterfactuals. Fundamental laws could help explain how we extend causal claims across different domains. And they could also help explain why autonomy assumptions are justified—why the operation of one causal mechanism is independent of another. For example, one could appeal to the fact that electromagnetic force declines with distance and the spatial separation of various parts of a system to justify why changing the charge on one part of a system won't affect the relations in another part. Woodward does appeal to fundamental laws to explain how we evaluate counterfactuals with physically impossible antecedents (2003, pp. 129–130). But perhaps fundamental laws have wider significance for interventionist accounts. This brings us to consider the third desideratum: relating causation to fundamental laws.

Even though appealing to fundamental laws would help interventionist accounts, these accounts don't do well at satisfying the third desideratum. Woodward explicitly denies that 'there must be a law of nature "underlying" or "backing" all true causal claims' (2003, p. 147). While he allows that there may be 'underlying' laws of nature, he takes it that this would be a contingent empirical discovery, and should not be part of an account of causation. Fundamental physical laws only feature in Woodward's account as generalisations that are *invariant* across a wider than usual range of interventions and background conditions—invariant in the sense that even if background conditions change, the generalisation still holds. Fundamental laws are the most invariant generalisations (2003, p.

240).[87] For example, 'we may think of the gravitational inverse square law as codifying information about exactly how manipulations that change the distance between two masses or the magnitudes of the masses themselves will change the gravitational force they exert on each other' (Woodward 2003, p. 67). This functional relation will be relatively stable even if other forces or bodies are introduced to the system. So it counts as a fundamental law.

But this approach does not adequately relate causation to fundamental laws. Firstly, it doesn't capture one of the key features of fundamental laws—their universality. The invariance of a relation is always taken to be relative to a particular system (2003, p. 251). If 'particular system' is significant at all, fundamental laws can't be both universal and identified with a single invariant generalisation. The fact that a law covers many or all such systems is in fact irrelevant to its being a generalisation that is highly invariant in any particular system. Secondly, even if fundamental laws can be adequately characterised as invariant generalisations, this approach doesn't relate fundamental laws to causal relations in non-fundamental systems. This is because all interventionist counterfactuals are relativised to a particular set of variables or causal mechanisms, and there is no attempt to explain how we can move between one system and another, or how generalisations that hold in one system can be used to give counterfactuals that hold in another. Even if fundamental laws could be characterised as invariant generalisations, this wouldn't secure any connection between higher-level interventionist counterfactuals and fundamental laws.

[87] Pearl takes fundamental laws to be completely invariant, even though their relevance to any given causal mechanism can change (2000, p. 210). Woodard takes fundamental laws to be only *relatively* invariant. But it's consistent with Woodward's view that the fundamental laws are completely invariant. The examples of only relatively invariant 'fundamental laws' from general relativity and quantum mechanics that Woodward gives aren't in fact candidate fundamental laws.

Nor does appealing to Pearl's primitive causal mechanisms help. The problem simply becomes one of relating or reconciling primitive causal mechanisms with fundamental laws, a problem Pearl does not address, except with regard to causal asymmetry (see chapter 5, section 5.1a). Some of these concerns might be less pressing if interventionist accounts at least secured *consistency* between fundamental laws and causation. But they don't. Some of these counterfactuals are explicitly taken to involve nomologically or physically impossible interventions (Woodward 2003, pp. 129–30).

Overall, interventionist accounts from Pearl and Woodward don't adequately answer the three desiderata on accounts of causation. They don't give us adequate insight into why we should care about causation, how it's accessible to us, and how causation relates to fundamental laws. Interventionist accounts are still interesting and explanatory, and do useful work relating various types of causal relations and counterfactuals to one another. These accounts succeed at many of their own aims. But they don't provide the kind of philosophical justification and elucidation of causation that could answer Russell's challenge.

### 4.3 Reductive Counterfactual Accounts

Reductive counterfactual accounts attempt to reduce causal relations to counterfactuals that can be evaluated without reference to causation. They aim to reduce causation to non-causal regularities (such as fundamental laws) and contingent features, using counterfactuals as a half-way step. So they may provide what interventionist accounts don't—a relation between causation and fundamental laws and a way into the causal framework. Both interventionist and reductive accounts relate causation to counterfactuals. But only reductive accounts aim to eliminate reference to causation altogether. Perhaps they can do better at satisfying the desiderata?

### 4.3a Lewis' Counterfactual Account

I'll take David Lewis' well-developed and popular account to be paradigmatic of reductive counterfactual accounts (1973a; 1973b; 1979a; 1979b; 1986).[88] Under Lewis' account, *c* *counterfactually depends* on *b* if and only if were *b* to occur, *c* would occur and if *b* were not to occur, *c* would not occur, where *b* and *c* are two distinct particular events. Counterfactual dependence is sufficient for causation, but not necessary. What is necessary is a chain of counterfactual dependencies between events that connects the two events in question. There need not be direct counterfactual dependence of one event on the other.

Lewis takes the causal relata to be events, a standard view influentially defended by Davidson (1980). In contrast, as we saw, interventionists like Woodward take the causal relata to be 'variables, or more precisely…changes in the values of variables' (2003, p. 112, emphasis removed). The values these variables take need not represent events. Lewis' account is also token-level and concerns particular events, rather than event-types (1973a, p. 558). Interventionist accounts are primarily type-level. My own view is closer to interventionist accounts on both these points. I'll later take causation to be primarily type-level and the relata to be states of affairs, represented by propositions. But for the moment I will simply evaluate Lewis' account as a theory of causation between particular token events.

[88] In the next chapter I also discuss statistical-mechanical accounts that derive from Lewis'. I don't discuss Lewis' later analysis of causation in terms of influence (2000).

Under Lewis' account whether *b* counterfactually depends on *c* depends on the truth of two counterfactuals: were *b* to occur, *c* would occur and if *b* were not to occur, *c* would not occur. Lewis relies on a comparative similarity ordering between possible worlds to evaluate these counterfactuals. A counterfactual is true (at the actual world) just in case either (1) there are no possible worlds in which the antecedent is true or (2) among worlds in which the antecedent is true, the consequent is true in at least one world closer to the actual world than any in which the consequent is false. By stipulation, no world is closer to the actual world than the actual world. So one of the counterfactuals we need to evaluate will be true or false merely in virtue of how the actual world is. For example, if both *b* and *c* occur, the counterfactual if *b* were to occur *c* would occur is trivially true. To evaluate the other counterfactual, we need to consider worlds where *b* does not occur, and determine whether *c* occurs. If a world where *c* does not occur is closer than any in which *c* does occur, the second counterfactual is true and *c* counterfactually and causally depends on *b*.

To evaluate non-trivial counterfactuals, we need to look to the comparative similarity ordering between worlds. Lewis takes this ordering to be a metaphysical primitive. However, in response to concerns from Fine (1975) and Bennett (1974), he identifies standards by which similarity is to be evaluated. Two of these standards are formal stipulations: each world is closest to itself, and any two worlds are comparable for closeness (still allowing for ties) (1973a, p. 560). The remaining standards for evaluating similarity are as follows, in order of importance (1979a, p. 472):

(1) …avoid big, widespread, diverse violations of law.

(2) …maximize the spatio-temporal region throughout which perfect match of

particular fact prevails.

(3) …avoid even small, localized simple violations of law.

(4) It is of little or no importance secure approximate similarity of particular fact, even in matters that concern us greatly.

Lewis intends these standard to pick out nearby worlds with a certain characteristic *structure*—one that hopefully allows the counterfactuals to deliver causal relations that correspond to our ordinary causal intuitions. For example, consider the counterfactual: if Nixon had pressed the button, a nuclear holocaust would have occurred. Lewis describes the nearby world $w_1$ to the actual world ($w_0$) as follows (1979a, p. 468) (see figure 6):

> Until shortly before *t*, $w_1$ is exactly like $w_0$. The two match perfectly in every detail of particular fact, however minute. Shortly before *t*, however, the spatio-temporal region of perfect match comes to an end as $w_1$ and $w_0$ begin to diverge. The deterministic laws of $w_0$ are violated at $w_1$ in some simple, localized, inconspicuous way. A tiny miracle takes place. Perhaps a few extra neurons fire in some corner of Nixon's brain. As a result of this, Nixon presses the button. With no further miracles events take their lawful course and the two worlds $w_1$ and $w_0$ go their separate ways. The holocaust takes place. From that point on, at least so far as the surface of this planet is concerned, the two worlds are not even approximately similar in matters of particular fact.

More generally, the counterfactual antecedent is satisfied at $w_1$ because a small miracle (a localised violation of the laws of $w_0$) occurs just prior to time *t*, the time of the antecedent. Other than at this small miracle, the actual laws hold at all times in $w_1$. The events at *w1* are exactly the same as at $w_0$ prior to the miracle, securing a large area of perfect spatio-temporal match. But after the miracle, the histories of $w_0$ and $w_1$ diverge in radical ways.

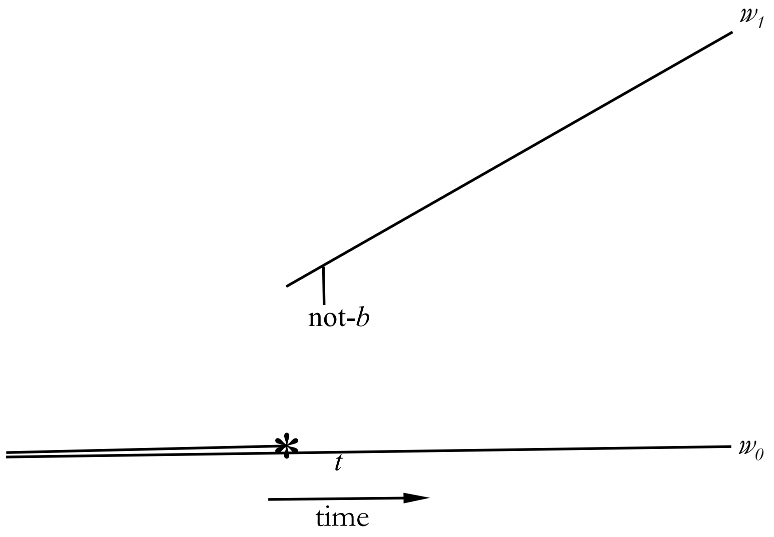


Figure 6: The differences between the actual world ($w_0$) and a nearby counterfactual world ($w_1$), given Lewis' standards of similarity. The worlds match perfectly until just prior to time *t*, when a small miracle (*) takes place, leading to the satisfaction of the antecedent (not-*b*) and massive divergences thereafter.

If the standards of similarity pick out worlds with this structure, Lewis' account may be able to distinguish between cases where correlations are due to effects of a common cause (figure 3, p. 125), or due to a direct causal relation (figure 2, p. 124). A perfect spatio-temporal match prior to the miracle means that the causes of *b* will be kept intact in the counterfactual world, while the effects of *b* may well be changed. If the causes of *b* are kept intact, as well as the laws (other than at the miracle), we may also be able to keep other effects of a common cause intact. These other effects will then not depend counterfactually on whether *b* occurs. *c* will only depend counterfactually on *b*, if *b* causes *c*. Or at least that is the hope.

Lewis' account shares feature in common with interventionist accounts: in evaluating whether *b* causes *c*, we consider what will happen in counterfactuals where *b* does not occur and see if this is correlated with changes in whether *c* occurs. But rather than considering an intervention that is itself a causal process, we have a small miracle that brings about *b* or not-*b*. And rather than relying on direct causal relations to evaluate counterfactuals, Lewis relies on explicit standards, including keeping the fundamental laws (mostly) fixed.

### 4.3b Evaluating Lewis' Account

A number of problems have been raised with Lewis' account. Adam Elga (2001) argues that Lewis' standards can't deliver nearby worlds with the asymmetric temporal structure Lewis intends. I consider these concerns in the next chapter (section 5.1b). Concerns are more commonly raised about cases of late pre-emption, where it seems there is causation but no chain of counterfactual dependence. The classic example is Billy and Suzy throwing rocks (Hall 2004). Suzy throws a rock and it smashes the bottle. Billy also throws a rock, but his rock arrives after Suzy's—too late to smash the bottle. If Suzy's rock hadn't shattered the bottle, Billy's would have. So the bottle smashing doesn't counterfactually depend on Suzy's throw. Yet it seems Suzy's throw *is* a cause of the bottle smashing. Moreover, unlike in early cases of pre-emption, there is no chain of counterfactual dependence between events connecting Suzy's throw with the bottle smashing. For any event on the chain connecting Suzy's throw with the bottle smashing, if that event had not occurred, the bottle still would have smashed, due to Billy's rock.

Paul and Hall (2013, ch. 3) give a useful overview of late pre-emption cases and the trouble they cause for counterfactual accounts. This trouble shouldn't be too surprising. As we saw with Russell's challenge, nothing less than information about every part of system is enough (given the laws) to guarantee an effect. Counterfactual accounts like Lewis' inherit this feature, insofar as they hold the laws fixed (except at the occurrence of the miracle). Changes anywhere in the system can contribute to whether an effect occurs. But if our intuitive judgments about causation are local (as Russell took them to be), we think changes at a distance from the putative cause (such as Billy's throw) shouldn't alter whether an event like Suzy's throw counts as a cause. And so we will get a clash between our (local) intuitive and (global) counterfactual judgments.

While my aim isn't to defend Lewis' account, my own account faces a similar concern, because I will also rely on counterfactuals. I won't give a full response to the problem, but point towards what I take to be the right approach: take our intuitive local causal judgments to be only defeasible guides to causation. We should explain why our local causal judgments work as well as they do, but not attempt to recover the full structure of our intuitive causal reasoning. For example, we might explain why we intuitively and reasonably judge Suzy's throw to be a cause of the bottle shattering by reference to the fact that our environment doesn't typically contain competing rock throwers like Billy. Given our typical environments, we can explain why we have intuition that Suzy's throw is a cause. But this isn't a reason to reject a global counterfactual account that implies Suzy's throw is not a cause (in the unusual circumstances where Billy is around). This strategy shares features in common with Pearl (2000, chs. 9–10), Hitchcock (2007), and the 'de facto dependence account' discussed by Paul and Hall (2013, p. 170). But, unlike some of these accounts, I don't aim to give a correct account of what conditions are required for particular causal claims to hold. I take it the most we should aim for is an explanation of why make the causal judgments we do in particular cases. I am sympathetic to suggestions by Hitchcock (2007, p. 512) and Woodward (2003, p. 85), that the patterns of counterfactual dependence are what are most central to causation, not particular causal judgments.

Leaving pre-emption cases aside, it's time to consider how Lewis' account fares with respect to the three desiderata. By analysing causation and counterfactuals partly in terms of laws, Lewis's account does better than interventionist accounts on the third desideratum: relating causation to fundamental laws. Fundamental laws are directly relevant for evaluating the counterfactuals that determine causal relations. Laws are kept intact at all points in the nearest counterfactual world, except at the time of the miracle. These miracles are a source of problems, however. For example, Lewis is led to accept that we are able to do things such that, if we were to do them, a generality that is a law would not be a law (1979b, p. 115). While he denies that this counts as being able to break the laws, this is partly a verbal stipulation. We would do better on both these points if the method of evaluating counterfactuals did not require miracles.

Are Lewisian miracles merely an optional part of the formalism? No. Miracles are closely tied to Lewis' rejection of 'backtracking counterfactuals': counterfactuals where, if the antecedent had been different, events causally upstream of the antecedent would have been different. Excluding backtracking counterfactuals is what allows counterfactual dependence to track causal dependence, under Lewis' account. For example, consider a case where low air pressure is a cause of a low barometer reading and a storm. Lewis rejects the relevant backtracking counterfactual, where the pressure depends on the reading of the barometer. Why?

> …[because] When something must give way to permit a higher reading [of the barometer], we find it less of a departure from actuality to hold the pressure fixed and sacrifice the accuracy, rather than vice versa. It is not hard to see why. The barometer, being more localized and more delicate than the weather, is more vulnerable to slight departures from actuality.
>
> (1973a, pp. 564–5)

Lewis recommends breaking the actual laws and introducing a small miracle that makes the barometer unreliable as an indicator of the pressure.[89] The alternative is to hold fixed the laws and the correct functioning of the barometer—and have a backtracking counterfactual, as in Bennett (1984). But then we'd have counterfactual dependence, without causal dependence. Miracles are introduced in order to rule out backtracking counterfactuals. And so miracles (or something like them) are essential to the formalism. While neo-Lewisian accounts of causation aim to avoid violations of the fundamental laws, they share the same rejection of back-tracking counterfactuals.

Given we have miracles, can we use them to make sense of an agent's (apparent) ability to intervene—and answer the second desideratum? The prospects don't look promising. Woodward's and Pearl's interventions were causal processes. Given agents might control such causal processes, there is some hope of connecting interventions to the activities or choices of agents. But Lewis' miracles are not actual occurrences. They are events that, were they actual, the laws would not be what they are. Nor are they events that we can bring about. Lewis takes our abilities to be tied to what we can *cause* to be the case (1979b). If he didn't make this restriction, our ability to bring about miracles would seem to imply that we can break the laws—a bad result, for Lewis.

We could attempt to understand an agent's apparent ability to intervene in terms of a power to bring about the counterfactual antecedents, and take the miracles to be an artefact of the formalism. But the problem with this move is that Lewis' account gives us no guidance for why we take it we *can* bring about these antecedents. Overall, the account doesn't help us

[89] Lewis' argument for this point is hard to make out and to reconcile with his standards for similarity (1979a). Requiring simply small departure overall would lead one to keep the future relatively intact.

understand why causation is accessible agents, leaving it mysterious why we take ourselves to be able to intervene in they way required to make use of causal relations.

Lewis' account also has trouble meeting the first desideratum—the pragmatic requirement to explain why causation should matter. Lewis claims that his account captures our intuitive judgements about particular cases of causation and counterfactual dependence. But it remains mysterious why a counterfactual account that operates in Lewis' terms should be useful to us. Why for example, does exact similarity matter so much, but approximate similarity not at all? Why hold the laws fixed, except for a small miracle? These apparently arbitrary-looking criteria leave it mysterious why a causal relation formulated in these terms tracks something we should care about—even if it delivers the right intuitive results. Bennett (1984), Horwich (1987, p. 172) and Woodward (2003, p. 137) raise similar concerns.[90]

These concerns are all the more pressing given that Lewis, at least in early work, emphasises the context sensitivity of counterfactuals and claims he is only giving one particular set of standards for their evaluation that may or may not be relevant in a given context (1973a, p. 565 n. 10). There's no attempt to say, for example, that a man who claims that if he had stepped out his window, he wouldn't have fallen to the ground, is doing anything other than using eccentric standards. But given the important roles counterfactuals and causal relations are meant to play in scientific and everyday decision-making and reasoning, the standards employed need to be justified.

[90] Lewis' failure to explain the usefulness of his criteria is particular a concern if we adopt causal decision theory, as Lewis does. Why should causation, understood in these terms, matter to rational decision-making? See Eells (1982, ch. 4) and Horwich (1987, ch. 11) for similar objections.

In some early work, Lewis seems to appeal to the intuitive plausibility of the standards: he talks of minimising 'departures from actuality' (1973a, p. 566) and of correctly tracking our 'opinions about comparative similarity' (1973b, p. 95). But he is later explicit that the similarity standards should not be defended on these grounds, but are instead reverse-engineered so as to pick out counterfactuals that deliver the right results: 'we must use what we know about counterfactuals to find out about the appropriate similarity relation—not the other way around' (1979a, pp. 466–7). Perhaps then the standards can be justified by what counterfactuals they pick out?

Here the account fares no better. It's unclear why the counterfactuals these standards pick out *are* the right ones: in particular, why ruling out back-tracking counterfactuals is warranted. On the face of it, it seems like changes to the present may well be counterfactually correlated with changes causally upstream. Lewis is explicit that his standards are tailored to rule out backtrackers (1986, postscript B to 1973a). But if the standards of similarity are justified *because* they rule out backtracking counterfactuals, we do no better than on an *a priori* strategy in understanding why ruling out backtracking is warranted. So we can't straightforwardly appeal to the counterfactuals to justify why the standards pick out a useful relation. We have too little understanding of why these counterfactuals are the right ones. Similarly if we justify the standards in terms of the causal judgments they lead to—we have too little deep understanding of why our intuitive causal judgements are useful.

Overall, while Lewis' account is more successful than interventionists in securing a connection between fundamental laws and causation, it has significant problems accounting for intervention and answering the pragmatic requirement. This is not to say that Lewis' account cannot reduce causation and counterfactual dependence. But, as with interventionist accounts, the accounts leaves interesting explanatory questions unanswered. We should consider what other kinds of philosophical elucidations of causation are available.

### 4.4 A Deliberative Account of Causation

In the remainder of the chapter, I put forward and defend a deliberative account of causation. This account is based, from the beginning, on what *will* be useful to deliberating agents. I'll argue that causal relations correspond to the evidential relations agents need when they decide on one thing in order to achieve another. Roughly, if deciding on a state of affairs of a given type, A, for the sake of a state of affairs of another type, B, is a good reason for believing that a state of affairs of type B will obtain, then A is a *cause* of B. More precisely, I'll defend the following biconditional, relating particular states of affairs picked out by types:

> *Evidential Biconditional:* A is a (type-level) cause of B *if and only if* an agent deciding on a state of affairs of type A in 'proper deliberation' for the sake of a state of affairs of type B would be good evidence of (a state of affairs of type) B obtaining.

This account of causation does well on the desiderata above: it explains why causation should matter to agents, how agents take it they can intervene, and it relates causation to fundamental laws. Because knowledge of causal structure is needed for good decision-making, we have a clear account of why causal relations matter.

Huw Price defends a related agent-based account, also arguing that causation should be understood by reference to agency (1991, 1992a, 1992b, and elsewhere). After presenting my account, I'll compare it to Price's. To prefigure, the main differences will be that my account employs a different model of deliberation, produces more objective causal relations, and is directly relevant for first-order metaphysical accounts.

### 4.4a From Deliberation to Causation

In chapter 2, I put forward four necessary and non-redundant conditions on an agent taking an option to be available in deliberation. Where **a** is a state of affairs represented by a proposition, an agent takes **a** to be available to decide on in deliberation if and only if:

(a) her deciding on **a** is a serious epistemic possibility for her;

(b) she is practically certain that if she decides on **a**, then **a**;

(e) she takes several options to be available;

(f) she takes her apparently-available options to be incompatible.

As noted, condition *b* involves a conditional. This may be a conditional belief, but a belief in a material conditional will suffice. These conditions were part of the epistemic model of deliberation. In chapter 2, I argued that this model explains why an agent takes different options to be available to decide on. I will now develop this model into a deliberative account of causation.

A first simplifying step is to combine conditions *e* and *f* into a single condition (*g*) on taking an option to be an available *alternative*, by definition, one of several apparently incompatible options.[91] Deliberation is then necessarily between options that are taken to be alternatives. An agent takes an option to be an available *alternative* in deliberation if and only if:

(g) she takes at least one other (apparently) incompatible option to be available.

A second step is to make the conditions more normatively demanding. We need to concern ourselves not merely with what agents take to be the case, but what they *should* take to the be case, based on good epistemic reasoning. This is one of the crucial moves that will make causation suitably objective. For the purposes of this account, I will assume a broadly evidentialist account of epistemic reasons. Under this account, whether an agent has reasons for a belief is settled by what evidence she has at the time, or, more generally, her epistemic grounds. (If one is uncomfortable at the term 'evidence' being used as an agent's epistemic grounds for her belief, the reader is welcome to substitute the latter.)

We also need a notion of 'good evidence'. This is evidence that is, in general, sufficient to justify an agent being practically certain of a state of affairs. While the account could be extended to cover lesser degrees of evidential support, the broader philosophical issues will be clearer if we deal only with evidence that justifies certainty.[92] For the purposes of this account, I adopt the following account of evidence. Firstly, I begin by taking the evidential relata to be states of affairs represented by propositions, rather than objects or events. This is keeping with the epistemic conditions on deliberation. Secondly, these sates of affairs need not be mental states, like having a belief or having knowledge. Evidence is the kind of thing

[91] This adjustment will simplify the derivations to come, but it also weakens the conditions slightly, since every option need not be incompatible with all others.

[92] Extending the account using a more general notion of evidential support fits nicely with a probability-raising approaches to causation, where A counts as a cause of B only if A raises the probability of B. A corresponding condition on intention might be that intending A requires taking one's intention to raise the probability of A. For details of this approach see Price (1991; 2012) and Mellor (1988).

to which others can have access—it has a public character. While mental states can count as evidence, they count as evidence equally for everyone, not just for the agent who is in the mental state. Thirdly, an agent's evidence is not relative to her background beliefs. For example, clouds count as evidence of rain for Tamsin even if she has no idea that these are rain-promising clouds. An agent may have evidence without believing or knowing she has evidence, or take herself to have evidence when she does not. Some of these failures may be based on the agent lacking or having false background beliefs. While we do have internalist concepts of evidence where evidence is relativised to the background beliefs of an agent, it is enough for my purposes that we have *a* concept that is external and non-relativized (and public) in the way I require. Fourthly, whether an agent *has* evidence can depend on what observations she has made—such as whether Tamsin has observed the clouds. Finally, the primary function of evidence is to justify belief. To this extent, evidence is normative, even if one can specify what evidential relations there are without appealing to norms.[93] Note that I haven't presumed evidence is always of prior states of affairs. This is crucial if the deliberative account is to *explain* the temporal asymmetry of causation, rather than presuppose it.

With this notion of evidence, changing the above epistemic conditions on deliberation into evidential conditions suggests the following three necessary and sufficient evidential conditions on what I call 'proper deliberation'. I will add one further condition later. An agent takes an option to be an *available alternative* in proper deliberation if and only if:

(h) her evidence does not settle her not deciding on **a**;

[93] It may seem evidence should be understood in terms of what is a reliable indicator of what, and so is a causal notion involving reliability. I address this concern in section 6.

(i) her deciding on **a** is good evidence of **a**;

(j) her evidence leaves open her deciding on at least one other incompatible option.

For example, say Tamsin is deliberating on taking her umbrella in order to stay dry. According to the conditions, if her deliberation is proper, her evidence can't already settle whether she will take her umbrella (or some other incompatible option). Yet her deciding to take her umbrella must be good evidence that she in fact will. Proper deliberation combines the epistemic model of deliberation with the requirement that an agent's beliefs are in line with her evidence. It marks out deliberations where agents are not only deliberatively reasonable, but also epistemic reasonable. As I argued in chapters 2 and 3, agents don't violate epistemic norms when they fail to have beliefs in deliberation about what they will do. So their evidence must leave open what they will do and how they will decide. Note that condition *j* refers to what options are *actually* incompatible. As in chapter 2, options can be incompatible for a variety of reasons—they may logically incompatible, or not jointly realisable for physical, technological or other reasons.

These conditions explain three features of deliberation. Firstly, they explain how the apparent openness of deliberation is *justified*. It is not simply that agents are uncertain of how they will decide; their evidence leaves open how they will decide. Secondly, the conditions explain why agents are justified in taking their decisions to be relevant for how the world goes. Their decisions *are* good evidence for their outcomes—it is not just that agents think they are. Combining these two features explains a third: why agents are justified in taking multiple incompatible states of affairs to be available to decide on. Their evidence leaves open several decisions, and these decisions are good evidence for incompatible states of affairs. This is how the deliberative account I go on to develop answers the second desideratum on accounts of causation: causal relations are 'accessible' to agents because agents are justified in taking incompatible states of affairs to be available to decide on.

So far these conditions capture how a decision is *evidence* for its result. The next step is to consider what's required for a decision to count as a *cause* of its result. Why might an evidential relation imply a causal relation? Consider condition *i* above: an agent properly deliberates when her deciding on **a** is good evidence of **a**. When all is going well, we expect this condition to hold for deliberating agents. Tamsin's deciding to take her umbrella *should* actually be good evidence she will, if she deliberates on taking it. While Tamsin may be prevented from taking her umbrella, she can't have good reason to suspect this, if she is to properly deliberate. Furthermore, we expect condition *i* to hold typically *because* an agent's deciding *causes* her to act as she's decided. So it shouldn't be surprising that a causal relation holds only if this evidential relation holds—even though I've explicated and defended the epistemic model of deliberation without appeal to causation.[94]

What about causal relations involving states of affairs that *aren't* decisions? I'll now argue that evidential conditions can capture these too. Above I argued for an important pragmatic requirement on accounts of causation—accounts should explain why causation matters to agents. Here's how we can satisfy this requirement. Agents need to know and care about whether their decision are actually good evidence of the outcomes they seek. If Tamsin is deliberating about whether to take her umbrella for the sake of staying dry, she needs to know and care about whether her deciding will actually be good evidence of her staying dry. If it is, she'll have made an appropriate decision. My claim is that a causal relation obtains between her taking her umbrella and her staying dry, just in case her decision to take her

[94] What about when a decision is evidence of a state it doesn't cause? I'll consider this problem briefly below, but my main treatment comes in chapter 5, section 5.3.

umbrella (for the sake of staying dry, in proper deliberation) *is* good evidence that she'll stay dry. If so, we have a straightforward account of why causal relations should matter—they matter to good decision-making.

As a first pass at capturing general causal relations, we might try the following condition. Where A, deciding A, and B are independent states of affairs of specified types, such that none is a constitutive part of any another, a state of affairs of type A is a type-level cause of a state of affairs of type B, if and only if:

(k) an agent deciding on A (in proper deliberation) is good evidence of B.

Tamsin's taking her umbrella is a cause of her staying dry, for example, if and only if her deciding to take her umbrella, in a proper deliberation, is good evidence of her staying dry.

Note that I am using an explicitly type-based characterisation of the evidential relata. As with interventionist accounts, the deliberative account is primarily concerned with general type-based regularities (smoking causes cancer) rather than with specific token instances (my friend's smoking caused his cancer). While there may be token-level causation, this is to be accounted for in terms of type-level causation.[95] The deliberative account is type-level because evidential relations derive from *kinds* of states (there being footprints) being general indicators of other kinds of states (there being someone at the window). While these evidential relations relate particular states of affairs (those footprints are evidence that Thimo was at the window), these hold because of general correlations (say between footprints of a certain size and someone of a certain height). Because both type-level causes

[95] For examples of how token-level causation may be derived from type-level causation, see Woodward (2003, p. 74ff.) and Pearl (2000 ch. 10).

and evidence work at a general level, the first relatum might be present without the second—a type-level cause may be present without its effect, or a piece of evidence may good evidence for an outcome, but the outcome not occur. And in both cases the second relatum may occur, by luck or chance, even if the first is not present.

What about the causal relata? In chapter 2, I took options in deliberation to be states of affairs represented by propositions. We decide on the way things are—states of affairs, picked out propositionally. Continuous with this, I take the causal relata also to be particular states of affairs of given types. But as far as I can tell, nothing in the evidential biconditional itself mandates this view. Like Woodward, I don't take the general form of my account to commit one to a precise view of the underlying nature of the causal relata.[96] One could take it that we deliberate on facts, events, or actions, and take causation to similarly relate facts, events or actions. What's important is that the options we decide on in deliberation take the same form as casual relata (or that they are easily convertible into each other). But for the sake of a simpler presentation, I will assume that the causal relata are particular states of affairs of given types, represented by propositions.

But condition *k* above still isn't quite right. Sometimes decisions are good evidence of states they don't cause, even when deliberation is proper. This is because deciding is often good evidence that the agent will take the *means* to their ends. For example, say Tamsin is properly deliberating on taking her umbrella, and must get to her umbrella by walking through the hall. If so, her deciding to take her umbrella is good evidence of her walking through the hall. But her taking her umbrella doesn't *cause* her to walk through the hall.

[96] I'm also sympathetic to Field's suggestion (2003, n. 6), that the difference between event-based views (Lewis 1973a; 1979b) and fact-based views (Bennett 1988; Mellor 1995) is largely terminological.

To deal with these cases, we need to appeal to a feature of instrumental reasoning, and take it that agents decide on one state 'for the sake of' another (Anscombe 1957, section 26).[97] Causal relations will only be implied when the cause is decided on 'for the sake' of the effect. For example, Tamsin's taking her umbrella counts as a cause of her staying dry if and only if her deciding to take her umbrella *for the sake of* staying dry (in proper deliberation) is good evidence of her staying dry. If Tamsin doesn't decide to take her umbrella for the sake of walking through the hall, then even if her deciding is also good evidence of her walking through the hall, the condition no longer implies her taking her umbrella causes her walking through the hall. With this suggestion, A is a type-level cause of B, if and only if:

(I) an agent deciding A for the sake of B (in proper deliberation) is good evidence of B.

There may be cases where we're tempted to say Tamsin *does* decide to take her umbrella for the sake of walking through the hall (perhaps she feels she needs the exercise). But in this case, Tamsin only decides in this way in order to get around irregularities and deficiencies in her abilities or motivations. If she really was just interested in the exercise, she should just decide to walk through the hall, and leave the umbrella out of it. Thinking about the umbrella is helping her to overcome weakness of will. Perhaps the additional goal of the umbrella helps her focus less on the exercise involved in getting it. While we do sometimes use decisions and deliberations in this way, causal relations shouldn't be tied to this aspect of

[97] A notion of 'basic action' can do similar work. Basic actions are those performed 'directly'—like waving an arm or moving a finger. They are standardly defined as actions that aren't done by doing other things—as I might signal a ship by waving my arms, or press a button by moving a finger (McCann, 1975). The claim would be that A is only a cause of B if A is a basic action. But while much about basic action is controversial, the concept of basic action is widely taken to be bound up with concepts from instrumental reasoning (see McCann 1975; Lavin 2013 and references therein). So we don't avoid appealing concepts from instrumental reasoning by appealing to basic actions.

deliberation. They're tied to the instrumental side, when we reasonably decide on one thing in order to achieve another.[98]

Another concern with this appeal to instrumental reasoning is that it imports too much causal structure, threatening to make the purported explanation of causation circular. It might seem that we properly decide on A for the sake of B only because A is a cause of B. In reply, note that not all forms of 'for the sake of' are causal—there is a more generic notion to appeal to. Annike might decide to go boating for the sake of spending a pleasant afternoon, or for the sake of some decision being made. In neither of these cases is there a direct causal relation involved—rather the boating constitutes a pleasant afternoon, and the decision to go boating constitutes a decision. The account does not imply such relations are causal, since the states of affairs are not independent of one another. But these examples demonstrate that there is a more generic notion of 'for the sake of' to appeal to.

I mentioned above that there was to be one further condition on proper deliberation. Now that we've introduced a notion of 'for the sake of', we can include it. An agent properly deliberates on A for the sake of B only if:

(m) her evidence does not settle B.

If Tamsin is deliberating about whether to take her umbrella for the sake of staying dry, her evidence can't already settle her staying dry. She can't, for example, have already heard a

[98] Why does instrumental reasoning have this structure? Attempting to answer this question would surely require a dissertation in itself. But I suspect it's partly tied to the fact that our decisions are better evidence for certain states of affairs (those 'closer' at hand) and less good evidence for other (more 'distant') states of affairs. Considering this feature in full would require dealing with probabilistic evidence—something I've abstracted from in presenting the deliberative account.

weather report that the day will be sunny and dry. Why is this condition plausible? Given that Tamsin must be uncertain of her decision in order to properly deliberate, if she is certain of B, she must be certain of B independently of whether she decides on A. And if she is certain of B regardless of whether she decides on A, then there is no point deliberating and deciding on A for the sake of B. The obtaining of B is already settled, independently of her decision.[99]

There's one additional step needed to give the deliberative account of causation. So far I've only considered cases where there actually is an agent properly deliberating. But causal relations obtain even when agents aren't deliberating, aren't properly deliberating, or aren't around at all. How can we work out the causal relations in these cases? The solution is to introduce counterfactuals, and appeal to the evidential conditions that *would hold* in proper deliberation. I'll say more in the next section about how to evaluate such counterfactuals. With this addition, the *evidential biconditional* on causation is as follows, where A, deciding A, and B are independent types of states of affairs, represented by propositions.

> *Evidential Biconditional:* A is a (type-level) cause of B *if and only if* an agent deciding on a state of affairs of type A in 'proper deliberation', for the sake of a state of affairs of type B would be good evidence of (a state of affairs of type) B obtaining.

According to the evidential biconditional, causal relations are directly related to what evidential relations would hold, were an agent to be properly deliberating. According to the

[99] An alternative approach is to require that A only causes B if deciding on A is good evidence for B, and *not* deciding on A is *not* good evidence for B. This alternative is preferable for generalising the deliberative account to deal with probabilistic causation. But condition *m* is independently plausible, and appealing to it will later help highlight how the evidential relations aren't simply relative to the deliberator's perspective.

deliberative account, such a biconditional is sufficient to give us a rich and illuminating account of causation, in ways I'll detail in the remainder of the chapter.

### 4.4b A Worked Example

In this section, I'll show how the evidential biconditional works by using it to distinguish between two different causal structures. But, before I do, let me note a problem that seems to arise with use of counterfactuals. It might seem that evaluating decision-counterfactuals will require an account of counterfactuals: perhaps Pearl or Woodward's interventionist account or Lewis' reductive account. If so, the deliberative account will not provide an independent account of causation.

While I won't offer an explicit account of how to evaluate these decision-counterfactuals, I will offer some techniques that suggest how we can evaluate them without assuming an independent account of causation. In section 4.4d, I'll consider how fundamental laws play a role in evidential reasoning and in evaluating counterfactuals. In this section, I introduce a principle for evaluating decision-counterfactuals that deals with cases where, in the actual world, an agent is prevented from properly deliberating by having too much evidence. I'll then use the principle to show how the deliberative account can distinguish between two causal structures. In chapter 5, I provide a more general explanation of why the account gets the right result in common cause cases. But, for now, the principle is useful for showing how the biconditional works and why it can plausibly deliver causal relations. The principle is as follows:

*Principle for Evaluating Evidential Counterfactuals*: Agents retain the evidence they have of particular states of affairs obtaining, as far as is consistent with proper deliberation.

According to the principle, if an agent can't properly deliberate, because she has evidence that settles how she'll decide, and we want to determine causal structure, we do the following. We remove the minimal amount of evidence required for deliberation to be proper, and break what evidential relations we must, all while keeping as much evidence as we can of particular states occurring. We then use the evidential relations that remain to determine the causal structure. Why might this principle be plausible? This principle reflects the fact that deliberation can become more complicated, so that an agent's evidence while deliberating does not determine what decision she makes. If so, if deliberation is proper, her decision is no longer evidence for states she has independent evidence of while deliberating. So evidential correlations will be broken, rather than agents giving up evidence of particular states.

To see how this principle works, I'll use it to distinguish between two causal structures, where the causal structures are already known. Doing so is counter to the spirit of the approach, since eventually we want to use deliberative account to work out the causal structure. But already knowing the causal structures will be useful for seeing how the principle works. Consider an example I introduced above. Kyle has noticed that after celebrating birthdays (C), he tends to feel euphoric (E). But he's not sure whether this is because celebrating directly *causes* his euphoria (a common cause structure, figure 3, p. 125), or because his celebrating causes him to drink (D) and drinking causes his euphoria (figure 2, p. 124). He's not sure whether D is a cause of E.

The deliberative account can distinguish the two cases. To determine whether D is a cause of E, we need to consider an agent who can properly deliberate and decide on D, and see if his deciding on D for the sake of E is good evidence of E. But say Kyle is not in a position to properly deliberate on D because he has too much evidence. He knows that he drinks every birthday he celebrates (and not otherwise). So merely knowing it's a birthday, settles his drinking. This is where the principle for evaluating counterfactuals comes in. It tells us what evidential correlations we need to break, to have a case of proper deliberation. The principle tells us that Kyle should keep his evidence that he's celebrating a birthday (C). But C should no longer be good evidence of D. This connection has to be broken if Kyle is to properly deliberate on D. Were Kyle to be properly deliberating, he would be certain he's celebrating a birthday, but justifiably uncertain about whether he'll drink or not. In this counterfactual case, is his deciding to drink (for the sake of feeling euphoric) good evidence of his feeling euphoric? If it is, D is a cause of E, suggesting the causal structure of the original case is as represented in figure 2 (p. 124). The causal structure in the counterfactual case can be represented as follows (figure 7):

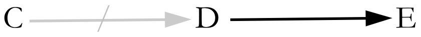


Figure 7: The causal structure that would obtain, were Kyle to be properly deliberating, given the causal structure in figure 2.

However, if Kyle's drinking is no longer good evidence of feeling euphoric when he properly deliberates, this suggests the causal structure in figure 3 (p. 125). Kyle's deciding on drinking (and drinking) is only correlated with euphoria when his celebrating settles his drinking. So if the evidential relation between C and D is broken, as it must be if Kyle is to properly deliberate on D, deciding on D is no longer evidence of E. In this case, D is not a cause of E. The causal structure in the counterfactual case can be represented as follows (figure 8):

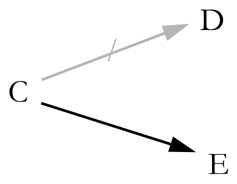


Figure 8: The causal structure that would obtain, were Kyle to be properly deliberating, given the causal structure in figure 3.

What allow the deliberative account to distinguish between the causal structures in figures 2 and 3 is the principle above and the fact that Kyle has evidence of C, the putative common cause. This evidence prevents his properly deliberating on D while C is evidence for D. In many cases agents will have evidence of common causes. In these cases, the evidential biconditional will correctly track the causal structure. But one may be worried that agents won't always have evidence of the common cause that prevents proper deliberation. This means an agent's deliberation can still satisfy the evidential constraints, and yet, it seems, his decisions may be evidence of the common causes. So-called 'Newcomb cases' are examples of this type. Since these concerns are closely bound up with temporal asymmetry of causation, I give my full treatment of them in chapter 5 (section 5.3).

### 4.4c The Objectivity of Causation

Before I evaluate the deliberative account using the three desiderata on accounts of causation, let me consider an objection that may already trouble the reader: the account won't deliver causal relations that are suitably objective. This kind of concern is commonly raised against agent-based accounts. Woodward (2003, p. 118 ff.), for example, claims that an agent-based account will burden us with an unnaturally subjective account of causation and imply a worrying form of mind-dependence. There are two ways in which the deliberative account secures the objectivity of causation. Firstly, the deliberative account is explicitly counterfactual—A is a cause of B if and only if, *were* an agent to decide A for the sake of B (in proper deliberation), A *would be* good evidence for B. Agents need not ever actually be in a position to deliberate on A. For this reason, causal relations do not depend on our practical or technological capacities to actually bring about A. Secondly, as stated, the account is concerned with the evidential relations that hold between a deliberating agent's decisions and other states of affairs, *no matter who is evaluating the evidence*. While the epistemic state of the deliberating agent plays an important role in picking which evidential relations are relevant, these are not evidential relations that hold merely from the agent's perspective—they are evidential relations that third-parties can observe and use.

The use of these more *objective* evidential relations is one way in which my account differs from Price's. Price defends a related agency theory of causation (1991; 1992a; 1992b; (with Menzies) 1993; 1996; 2007; (with Weslake) 2009; 2012; forthcoming). He argues that it is only when we take up the point of view of a deliberating agent, with its peculiar epistemic features, that causation appears in our world-view. Causal regularities are those correlations that hold 'from the free agent's distinctive point of view' (1991, p. 173) and that are 'assessed from the agent's distinctive epistemic perspective' (2012, p. 494). The idea is that the deliberating agent has a peculiar epistemic stance on the world, because she takes her decisions to be probabilistically independent of everything except their effects. So if we look at things from her perspective, there aren't probabilistic relations between her decisions now and their causes. The concept CAUSATION is in this sense 'perspectival'. In early work, Menzies and Price (1993) characterise this as a 'response-dependent' account of causation. In the same way that the concept RED depends on our responses to red objects, the concept CAUSATION depends on our agential engagements as agents. While in later work Price (forthcoming) distances himself from this original response-dependent formulation, he continues to maintain that understanding the role CAUSATION plays in our lives requires acknowledging its perspectival character.

But while I agree with Price that deliberation is crucial to understanding causation, this is not because the evidential connections are assessed only from the deliberator's perspective, or only hold for her: they are the kind of evidential relations that anyone can use. They are, in this important sense, objective. While I don't go as far as Mellor (1988) in taking causal relations to track *fundamental* temporally asymmetric probabilities, causal relations are more objective than Price supposes them to be. It is for this reason that I take my agent-based account to have important metaphysical consequences, while Price's does not. I take up these metaphysical and perspectival issues further in section 4.5.

### 4.4d Evaluating the Account

How does the evidential model of causation fare with respect to the three desiderata on accounts of causation? Regarding the second desideratum, explaining how causation is accessible to agents, the account does well. Without invoking primitive agent abilities, or explicitly causal notions, the account explains why agents are *justified* in taking incompatible options to be available to decide on in deliberation. They are justified because their decisions are good evidence for their results, and their evidence leaves open incompatible options.

Regarding the first desideratum, the pragmatic requirement to explain how causation is useful to agents, the account also fares well. Causation is useful to agents because it tracks the evidential relations agents use when deciding on one thing in order to achieve another. Because evidential reasoning is useful in such contexts, causation is too. Even theorists like Woodward who are inimical to deliberative approaches accept that correspondence with evidential relations (what would happen were an agent to act in a certain way) allows an account to straightforwardly satisfy the pragmatic requirement (2003, p. 31).

To see how the account fares with respect to the third desideratum, relating causation to fundamental laws, we need to return to a related concern: how do we go about evaluating the counterfactuals appealed to by the deliberative account? I introduced a principle above to deal with cases where an agent has too much evidence to properly deliberate, and we already know the causal structure. But what about cases where an agent has too little? What if an agent can't properly deliberate because her decisions wouldn't be evidence for the relevant states? And what about cases where we don't already know the causal structure? Are there more general guidelines for what evidence an agent would have, were she to properly deliberate? For example, say a molecule of hydrogen gas moving in such and such a way causes the movement of some other. This seems compatible with thinking that no actual agent ever had, or could have, the right kind of evidence so as to properly deliberate on how a particular hydrogen molecule moves. How do we then tell what evidential relations there would be in proper deliberation, without appealing to causal relations? Or how do we know that volcanoes erupting (von Wright 1971) or earthquakes occurring (Menzies and Price 1993) are the causes of their effects, if no agent was ever in a position to properly deliberate on volcanoes erupting or earthquakes occurring?

We need to be careful about the nature of the concern here. The concern isn't that actual agents lack the required evidence. Recall, nothing in the deliberative account requires agents to *actually* be in a position to properly deliberate, in order for states of affairs to count as causes. The crucial requirement is that *were* an agent to properly deliberate, an evidential relation would hold. Furthermore, the concern isn't that the account appeals to counterfactual evidence. There isn't anything unusual in talking of what evidential relations would hold, even if the relevant states never or could never obtain. For example, it may be true that *were* an asteroid to appear in a particular quadrant of the sky this *would be* evidence that it would impact the earth—even if no asteroid ever appears there, or even if no asteroids *could* appear there (due to the orbits of the other planets, for example). The concern is that we need some guidelines for evaluating such evidential counterfactuals.

We might simply claim we have easy access to facts about what would be evidence for what, were an agent to deliberate, by simply generalising from cases where we do deliberate, to cases where we do not and cannot. This is one of the responses made by von Wright (1971, p. 72) and the response preferred by Price (forthcoming). But even if such counterfactuals do turn out to have reasonably well-defined truth-values, we shouldn't help ourselves to them too easily. Otherwise it's unclear why we couldn't simply help ourselves to interventionist counterfactuals to begin with (as Woodward does) or other cause-involving counterfactuals. We need to know why generalising from the actual to the impossible is easier along agency lines (using deliberation), rather than along interventionist lines (using causal relations). Price (forthcoming) in fact presents his defence as a *tu quoque* to Woodward's objections (2003, p. 125): Woodward must himself rely on generalising from cases where we do intervene to cases where we do not and cannot. But while an agency account may do no worse than Woodward's on this point, this defence is insufficient. Evaluating these evidential counterfactuals remains a challenge.

One way to address this challenge is to see what might license the generalising moves from cases of actual deliberation. Although this is not his preferred response, Price (forthcoming) distinguishes two ways this generalising might be licensed: either *directly* from counterfactuals that hold in one system to counterfactuals that hold in another using inference principles of science (such as symmetry principles) or *indirectly* going via similarities in non-causal and non-counterfactual features of the systems. I am sceptical that we have inference procedures that are licensed directly and move us between the complex counterfactuals that hold in individual systems. Rather, there are physical symmetries in the fundamental laws and the systems that explain how we infer from counterfactuals in one system to counterfactuals in another. Even if we do use symmetries in counterfactual reasoning, and don't infer by appealing to fundamental laws, our ability to make the higher-level inferences is justified ultimately in terms of there being symmetries at the fundamental level. Even if we make the inferences directly, they are licensed indirectly. Spatial, temporal, and other symmetries in the fundamental laws are what explain why higher-level symmetry-based inferences between systems are licensed. If so, evidential counterfactuals should be related to fundamental laws. There are now two reasons for the deliberative account to relate causation to fundamental laws: Russell's initial challenge and because fundamental laws underwrite evidential counterfactuals.

If fundamental laws are used in the evaluation of evidential counterfactuals, then we can use the universality of fundamental laws, combined with physical similarities between different systems, to evaluate counterfactuals where agents have insufficient evidence to properly deliberate. For example, consider the colliding hydrogen gas molecules again. In kinetic molecular theory, we assume that gas particles collide elastically, conserving total kinetic energy. By idealising, we can then model the behaviour of the gas particles using more familiar everyday objects that also collide (approximately) elastically—like billiard balls. And agents do have experience of intervening on the motions of billiard balls. Because of physical similarities between systems of gas molecules and systems of billiard balls, and the universality of the fundamental laws that govern their behaviour (roughly Newton's second law), we can consider what evidential relations would hold, were an agent to properly deliberate on the motion of a gas molecule.

Is there a more straightforward way to derive evidential counterfactuals from fundamental laws? One approach would be to directly derive evidential counterfactuals from fundamental laws and contingent states. We'll see elements of this approach in the next chapter. But while such a derivation might be in principle available, it is not a story I expect can be told in detail anytime soon. And until such a story is on offer, the suspicion from non-reductivists like Woodward and Pearl will be that any such explanation must go via causal relations—so we won't have an independent account of causation. I don't have an account of what contingent features should be held fixed when evaluating evidential counterfactuals and working out what evidential relations will hold in counterfactual cases. I suspect this will be a highly context-dependent affair. It's more important for my purposes that fundamental laws are held fixed. This is enough to relate causation, under the deliberative account, to fundamental laws, and to give us some guide to evaluating evidential counterfactuals. What I'll argue now is that the evidential features of universal laws suggest they should be held fixed when evaluating evidential counterfactuals.[100]

Fundamental laws are, amongst other things, pervasive exceptionless regularities, precisely the kind of relations that we would expect to be good for evidential reasoning (aside from whatever modal features they have). One way of bringing out this fact is to look to Humean analyses of lawhood, such as Lewis' Best Systems Account. Under Lewis' account, fundamental laws are axioms of a system that best summarises information about what events occur, and does so in a way that balances strength (how much information about an event can be inferred) and simplicity (how many and how complicated the axioms are) (1973b, pp. 73–4; 1983, pp. 365–8; 1986a). Whether or not one accepts Lewis' account as a reductive analysis of lawhood, it has generally been accepted as identifying epistemological features of fundamental laws at our world—when concerns are raised, they are raised on other grounds (Carroll 1994; Roberts 2008). If fundamental laws are universal exceptionless regularities, this is a reason for holding them fixed when evaluating evidential counterfactuals. They are regularities we reasonably expect to hold across different systems in the actual world.

Any account of causation that makes use of fundamental laws in evaluating counterfactuals in a way that is meant to be philosophically edifying owes an account of why laws should be held fixed. As I've argued, answering the pragmatic requirements requires justifying why the standards appealed to are the right ones for giving useful causal relations. The deliberative account can appeal directly to the important epistemic role of fundamental laws in order to

[100] Although of course not when we reason evidentially about what the fundamental laws might be.

explain why laws should be held fixed. And so it secures a strong connection between evidential counterfactuals, causation and fundamental laws, satisfying the third desideratum.

If the deliberative account can appeal to fundamental laws as part of explaining why certain evidential counterfactuals hold, this suggests the following explanatory ordering—the layer cake picture:

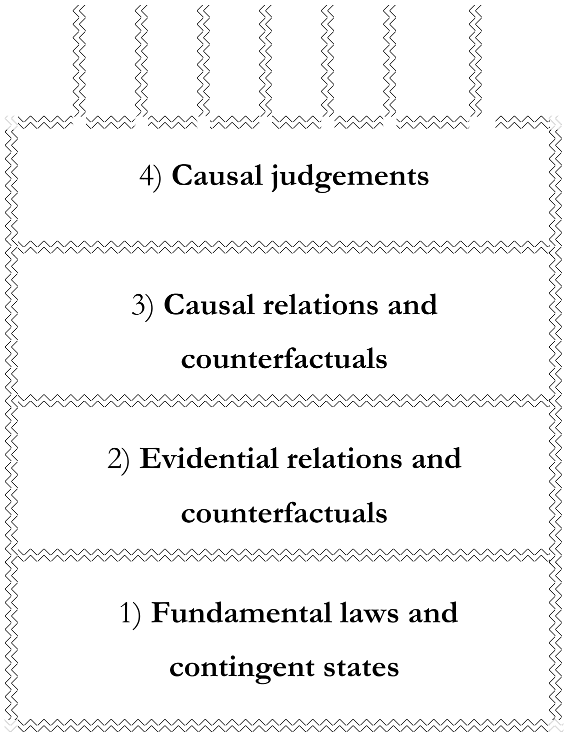

4) **Causal judgements**
3) **Causal relations and counterfactuals**
2) **Evidential relations and counterfactuals**
1) **Fundamental laws and contingent states**

Figure 9: The layer cake picture, showing the explanatory order that the deliberative account of causation supports.

At layer 1 there are fundamental laws and contingent states of affairs. No causal or counterfactual relations appear at this level, beyond what may be implicit in fundamental laws. These laws and states of affairs are then used to account for the kinds of evidential relations that hold, or would hold in counterfactual cases (layer 2). These evidential relations capture how localised states of affairs are good evidence for others. The evidential counterfactuals are roughly analogous to Price and Weslake's 'hypothetical conditionals'

(2009, p. 437) and mirror the kind of inferential uses of causal networks explored in Pearl (2000, chapter 2). They operate in both temporal directions. These evidential counterfactuals are then used to account for causal counterfactuals, using the evidential biconditional (layer 3). Causal relations correspond to the evidential relations that would hold for properly deliberating agents. We don't just begin with cause-involving counterfactuals, as on the interventionist account, or move from fundamental laws to causal counterfactuals directly, as one Lewis' account. There are further steps between causal counterfactuals and particular causal judgments (layer 4), such as 'the striking of the match caused it to light'. Recall, the evidential model is type-level—a further step is required to give token-level causal claims. In addition, the deliberative account does not mark a distinction between causes and causal conditions. Distinguishing these requires introducing further pragmatic considerations, such as relevance, which I have not explored here. And, as noted with regards to Suzy's rock-throwing, I don't take delivering causal judgments to be crucial work for an account of causation.

This layer cake picture fits with a reductionist program that accounts for causal claims and counterfactuals in terms of evidential relations and counterfactuals, and which reduces these to fundamental laws. But a more conciliatory approach suggests that the important connections here are explanatory and justificatory, rather than metaphysically reductive. This is the approach I recommend in the following section.

### 4.5 Metaphysical and Explanatory Upshots

If we accept the evidential biconditional and the deliberative account as providing necessary and sufficient conditions on causation, what follows for the metaphysics of causation? I will argue that the evidential biconditional is compatible with a range of metaphysical views. It also acts as an important constraint on first-order metaphysical accounts, one particularly relevant for reductive realist accounts.

As mentioned, Price defends a related agent-based account. But he takes his account *not* to have important metaphysical upshots. Price most commonly presents his account as a 'perspectival' account of the concept CAUSATION: understanding the concept requires appealing to the evidential relations that hold from the deliberator's point of view (2007, p. 254). Causation is 'conceptually dependent on human capacities and responses' (with Menzies 1993, p. 192 n. 12). While this might suggest a metaphysical thesis about the nature of causation, a kind of response-dependence account that I consider below, more often Price eschews talk of metaphysics altogether (2007, p. 255). If we take Price to be giving a conceptual account, there are still a number of forms the conceptual dependence might take. Price might be giving an 'analysis of [the concept] causation' (1991, p. 159), conditions for understanding the concept (1993, p. 201), an account of the function of the concept in our lives, or a genealogy of how we acquire the concept, either as individuals or as a species (1992a, p. 514).[101] Another option is to take the account to give an epistemology of causation—how we come to know the concept applies—as in Mellor (1988). More recently, Price is most interested in giving a genealogy and functional account of the concept, which explains the role of causal concepts and how we came by them. Price is explicit that 'the agency theory…should be seen as what I have sometimes called philosophical anthropology: the task of explaining why creatures in our situation come to speak and think in certain

[101] Price isn't explicit about how the functional and genealogical stories relate, but presumably each is meant to inform the other. Price is also not clear how much of the genealogy should be a scientific account of how the concept arises, and how much is a conceptual genealogy in the style of Bernard Williams (2002). For another example of the genealogy he has in mind, see Price (1988).

ways'—'we are not concerned with the project of providing a reductive *analysis* of the concept of causation, but rather with the anthropological project of explaining its genealogy and use' (forthcoming, emphasis in original).

But while I have no objection to using the deliberative account to give a conceptual genealogy in Price's style, this leaves other options, particularly metaphysical and reductivist options, underexplored. By tying a deliberative approach too closely to philosophical anthropology, we miss the relevance of deliberation and evidential relations for more realist approaches. In particular, non-perspectival first-order accounts can still use evidential features of causation to *explain* why we should care about the relations they pick out as causal. Price and Weslake are mistaken to claim that realists about causation have no response to the pragmatic requirement beyond 'a blunt appeal to intuition' (2009, p. 428).[102] In addition, by focussing on the pragmatic requirement alone, Price and Weslake miss the possibility that other requirements may be important in giving an account of causation. Satisfying other desiderata may be consistent with satisfying the pragmatic requirement well enough. One particularly important desideratum, I have argued, is connecting causation to fundamental laws. One of the advantages of reductive approaches it that they help satisfy this desideratum.

As far as there is a metaphysical account of causation to be found in Price's work, it is a 'response-dependent' one: 'causal relations are…mind (or more particularly agent) dependent, just as secondary qualities are sensory agent dependent' (1991, p. 173). Causal relations depend on us. But because we're unaware of this dependence, we naturally *project*

[102] In other places Price accepts that certain realists (who takes the perspectival account to be part of the role-fixing for the concept) can help themselves to his account of the concept's usefulness (2007, p. 287 n. 28).

onto the world relations that are partly determined by our own contingent perspectives, and take causation to be an objective feature of the world (1991, p. 173; 2007, p. 254). A response-dependent account of causation, combined with this projectivist semantics, might appear to have the best of both worlds—capturing the apparent objectivity of our causal language and while still explaining the relevance of causation for us.

But a deliberative account of causation is compatible with other metaphysical approaches. Figure 10 depicts some of these. Laying out these options helps reveal the flexibility of the deliberative account, and the wide variety of metaphysical positions it can support. Note that we might also adopt the accounts below as accounts of the *concept* CAUSATION—for the most part, I will take this option to be implicit in what follows.

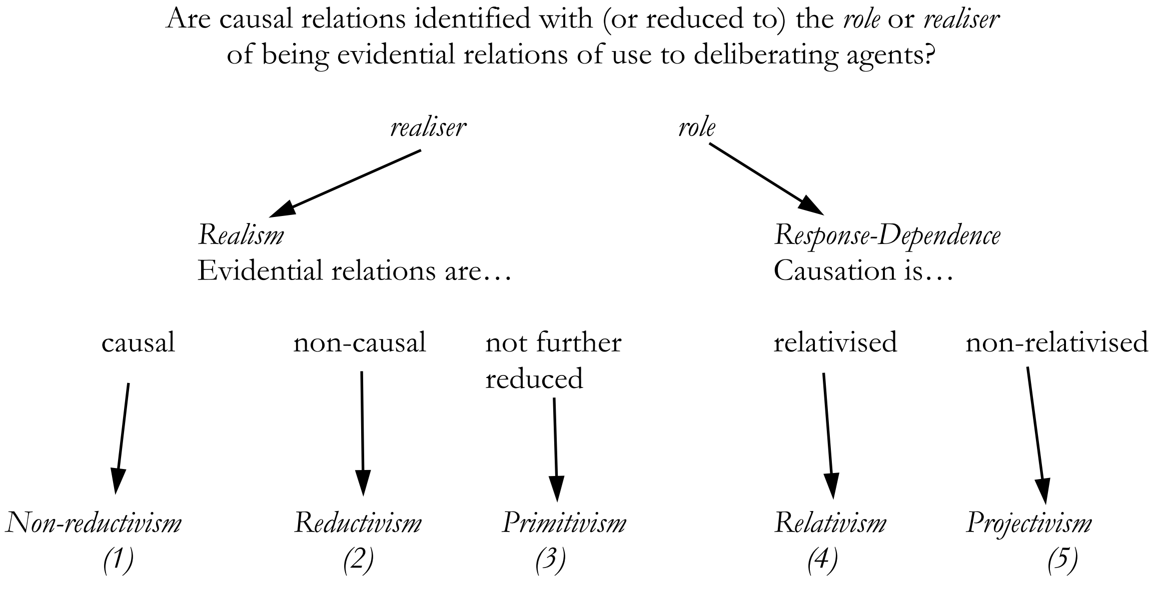


Figure 10: Metaphysical positions one can adopt towards causation, if one accepts the evidential biconditional as giving necessary and sufficient conditions on causation.

Picking up on a distinction in the response-dependence literature, causation may be identified with its evidential *role* (to give response-dependence), or it may be identified with what *realises* this role in the actual world (to give realism). In conceptual terms, the concept may be analysed in terms of its role, and so refer non-rigidly, picking out different realisers at different worlds. Or it may be analysed as the realiser, and so rigidly refer to whatever satisfies the role at our world, picking out the same realiser at different possible worlds. While all the positions in figure 10 agree on the extension of the concept causation in the actual world, they may disagree on its extension at other possible worlds.

On the right hand side of figure 10 there are two response-dependent views that directly relate causation to agency. What I'll call a *projectivist view (5)* rigidifies on *our* perspective as agents: causal relations are what would be the appropriate evidential relations *for us*, were *we* in the relevant possible world. A *relativist view (4)* does not rigidify on our perspective. Instead, causal relations are the appropriate evidential relations *for agents in the relevant possible worlds*. There are other options, but these are good representatives of response-dependent views. These views are interesting philosophically, and provide resources for identifying causal relations at possible worlds that have quite different metaphysical or nomic structures from our own. But they provide few resources for connecting causation to fundamental laws and so won't by themselves satisfy the third desiderata on accounts of causation.

On the left hand side of figure 10 there are three realist views. These identify causal relations as what realises the evidential role for actual agents in the actual world. Causal relations at other possible worlds can then be identified without considering their relevance for agents. According to these realists, while appealing to deliberating agents may play some role in picking out causal relations, causal relations exist and have their character independently of agents.

*Non-reductivist views (1)* take the realisers to be causal relations. These views do not attempt to reduce causal relations to anything else via the evidential biconditional. If interventionists like Woodward accepted the evidential biconditional, they would fall into this camp, since the realisers of the biconditional are simply more causal relations. Views that take evidential relations to be given by causal dispositions or causal propensities also fall into this camp (Mellor 1988). While such views can be defended using the evidential biconditional, they don't do well at relating causation to fundamental laws or explaining the temporal asymmetry of causation (as I'll argue in chapter 5).

*Reductivist views (2)* reduce causal relations to evidential relations (and counterfactuals) and in turn reduce evidential relations to non-causal relations. If they accepted the evidential biconditional, Lewis and neo-Lewisians would fall into this camp. Evidential relations would then be an intermediate stage in the reduction of causation to non-causal relations. Taking this approach means we don't rule out backtracking counterfactuals by fiat, as Lewis does. In chapter 5, I'll further explore how the deliberative account can support reductivist accounts.

*Primitivist views (3)* take the causal relations to be evidential relations, but don't reduce evidential relations further. I have labelled these views 'primitivist' since they may take evidential relations to be metaphysical primitives. Some theorists who fall in this camp may also be sceptical or agnostic about metaphysical reduction. Views that take objective chance or non-causal propensities or dispositions as primitives also fall into this camp, if these capture objective evidential relations. Some subjectivist views about evidential relations also belong here, such as view that relativise evidential relations to the agents at a world—producing a relativized causal relation. Further work would have to be done to relate causation to fundamental laws on any of these primitivist accounts. Perhaps one could reduce fundamental laws to evidential relations, but this is not an approach I will explore.

The evidential biconditional is compatible with all these metaphysical accounts and plays several roles with respect to them. Firstly, the biconditional acts as a constraint—whatever relation is to count as causation must satisfy the evidential biconditional for agents in the actual world. Secondly, the biconditional can be used to *justify* why a given account picks out a relation that deserves to be called causal. The evidential biconditional can play this role, moreover, even if one rejects it as giving precise necessary and sufficient conditions on causation. One may, for example, deny the biconditional is sufficient by requiring causation to be spatiotemporally continuous (Salmon 1984), transmit marks (Salmon 1984) or transfer conserved quantities (Dowe 1992a). In these cases, the evidential biconditional may still be important for explaining why causation is useful to agents in normal deliberative contexts. Accounts may differ in the extension of causation for other reasons, and also still be broadly justified by this evidential route. The pragmatic requirement requires causation to be useful in general. This is compatible with it giving 'bad advice' in particular instances—whether bad advice is determined evidentially or by other means. This point is particular relevant when we move to counterfactual situations. Accounts may disagree over what situations or with what degree of generality the norms of causation must deliver the 'right' results, and yet be well justified by this deliberative route.

Thirdly, the evidential biconditional can also elucidate the conceptual connection between causation and evidence—no matter what metaphysics of causation one adopts. The biconditional might not capture all aspects of our ordinary notion of causation. Our ordinary notion may, for example, build in a commitment to causes and effects being spatiotemporally continuous. But scientifically-informed philosophical reflection on causation, of the type many have pursued, may lead us to revise this commitment or at least to develop a new and more philosophically developed concept of causation.

The evidential biconditional can play these roles for all the above metaphysical approaches—including reductive accounts. But while response-dependent and even interventionist views have been considered as instantiations of the evidential model before, realist views like reductivism have not been similarly explored. Part of the reason seems to be scepticism over causation being a local fundamental relation (Price 1991, p. 160). But this attitude fails to distinguish reductive from other realist views. Not all reductive views take causation to be local, and they deny causation is fundamental.

A particularly appealing reductive option is to take causation to be a non-local relation that reduces to laws and contingent features—the same features that explain features of deliberating agents. If we're interested in relating causation to fundamental laws, this type of reductive account is particularly useful. In the next chapter, I'll consider in further detail how statistical-mechanical accounts of causation of this form are compatible with the deliberative approach. Note that one can also adopt much of the structure of such a reductive view, without committing to laws being metaphysically more fundamental than causation. Reductive views can be a way of connecting relations, without implying metaphysical asymmetry. My own view is that metaphysical reductions are explanatory, and that explanation does not require dependence of the explanandum on the explanans: explanations unify aspects of the world. It is in this sense that I am most interested in reductive accounts.

Given that reductive accounts are likely to have the most trouble meeting the pragmatic requirement, it is particularly important that they can appeal to the evidential biconditional to justify their accounts. As we saw above, one of the main concerns with Lewis' view was that his conditions on causation appeared *ad hoc* and unilluminating. If a defender of a reductive view can explain why their analysis picks up on the evidential relations of use to deliberating agents, they can justify their account. Altogether, while a range of metaphysical views are compatible with the evidential biconditional, it is particularly important for reductive accounts.

### 4.6 Concerns with Circularity

I end this chapter by considering a major source of scepticism regarding agent-based accounts of causation: problems to do with circularity. Hasuman argues that an account that uses terms like 'manipulation' and 'agency' will be unilluminating and viciously circular, since these terms are themselves causal (1997; see also Woodward 2003, pp. 123ff). In this section, I'll argue that the deliberative account remains substantive and illuminating, despite these concerns. This is due in part to the particular roles the evidential biconditional plays under the deliberative account.

Consider the first role of the evidential biconditional: providing a constraint on first-order metaphysical accounts. Circularity concerns do not affect whether the evidential biconditional can play this role—constraints can be circular, and still function as constraints. Circularity concerns also do not directly affect the kind of philosophical anthropology Price pursues—one can still give an illuminating account of how we acquire a concept or what role it plays, even if we have to use to the concept. Explaining how people came by the concept PEOPLE, for example, may involve referring to people, without making the explanation unilluminating. But circularity concerns do bear on two other roles the evidential biconditional might play—providing conceptual insight into causation, and a justification that reductivists can appeal to. It is to these circularity-based concerns I now turn.

One concern regarding circularity was first raised in chapter 2. It seems that if characterising deliberation requires appealing to causal concepts, we can't appeal to deliberation to give a non-circular account of causation. This concern was met by characterising deliberation in *epistemic* terms. Because it is based in a epistemic model, the deliberative account remains illuminating. In addition, if the deliberative account is right, and we develop causal concepts to suit deliberation, it is no surprise that aspects of deliberation have come to seem causal as well. Finally, note that this concern regarding circularity would be more pressing, were one to defend a response-dependent account of causation, in which deliberation is metaphysically or conceptually more basic than causation. But I am not defending a response-dependent account. It is enough for my purposes that deliberation can be well characterised in epistemic terms.

My use of an epistemic model of deliberation is another way in which my account differs from Price's. I have been careful not to characterise deliberation in causal terms, even when considering how deliberation appears to agents 'from the inside'. For example, I have not followed Price in making informal statements of the view like the following: '*A* is a cause of a distinct event *B* just in case *bringing about A* would be an effective means by which a free agent could *bring about B*' (Menzies and Price 1993, p. 193, their emphasis). Nor have I required that agents 'see their own actions as "uncaused," at least in the midst of deliberation' (Price 2012, p. 536), or that agents think of their deliberation as 'issuing in rather depending on the ensuing actions' (Price and Weslake 2009, n. 25). Nor are agents required to treat their actions as ultimate contingencies. Even within the agent's point of view, we should avoid using causal concepts to characterise deliberation, or more serious circularity concerns arise.[103]

In chapter 3, I also argued against epistemic accounts of deliberation due to Velleman, Joyce and Ismael. I didn't reject them because they used causal notions. But it's worth noting that these approaches are also unsuitable for defending a deliberative account of causation—a move Price suggests (2012, p. 529–30). Velleman characterises decisions in terms of their causal properties—they are 'self-fulfilling beliefs' (1989a, 1989b). Joyce is a causal decision theorist who argues that requirements on rational deliberation cannot be characterised purely in evidential terms—for example, agents must have a *causal* ability to decide on options (2002, p. 88). Even Ismael, I argued, relies on an unexplained apparent ability on the part of the deliberator. Until we see how to characterise this ability in non-causal terms, the suspicion will be that all these accounts rely directly on causal features of deliberation.

Price does not use an epistemic characterisation of deliberation to defend his account from the charge of circularity. Instead he appeals to our direct experience of agency: 'we do not

[103] Hitchcock (1996) also appeals to the perspectival character of deliberation to defend a probabilistic account of causation. My concerns with Price's account of deliberation carry over to Hitchcock's.

need an explication of agency in terms of causality. Agency is something of which we all have direct experience' (1991, p. 173). Concepts like 'bringing about' can be introduced by ostension, since 'we all have direct personal experience of doing one thing and thence achieving another' (Menzies & Price 1993, pp. 194–5). But while we may have such direct experiences, this doesn't show that causal concepts are not bound up with them. Practical deliberation, under Price's account, may still depend on agents using causal concepts. An epistemic characterisation of deliberation does better, by helping account for causation without directly appealing to causal notions.

A second concern with circularity relates to the pragmatic requirement itself. I required accounts to explain how causation or causal knowledge is useful to agents, and USEFULNESS seems to be a causal concept—a useful relation is one that helps us achieve our ends. In one sense, this circularity is benign. The pragmatic requirement is a general constraint on accounts of causation. Because it is not part of any account, it does nothing to trivialise accounts that satisfy it. This source of circularity can also be seen as a positive feature, to the degree that it may account for some of the indeterminacy in our concept of causation—different accounts of causation may satisfy the pragmatic requirement equally well, by their own criteria for usefulness. But an account will do better if it answers the pragmatic requirement by doing more than simply presuming that the causal relation it posits is useful, by the standards of that causal relation (as Woodward does). The epistemic account satisfies this further goal, because it appeals to the usefulness of evidential relations, rather than the usefulness of causal relations.

A third concern with circularity has to do with evidential relations. It seems that accounting for why a certain state of affairs counts as evidence for another may require appealing to causation. This concern puts pressure on primitivist accounts that take evidential relations as fundamental, and on response-dependent accounts that appeal to evidential relations as part of reduction or analysis. But my main concern here is that it also puts pressure on whether the evidential biconditional can be conceptually illuminating. In reply, note that even if we can and should give causal explanations for evidential relations, this is not enough to show that appealing to evidential relations delivers a trivial or uninteresting account of causation. We expect almost any observed regularity to be amenable to causal explanation—this has become a presupposition of much of science. But this does not mean that evidential relations cannot be understood independently of causation. Originally I introduced evidential reasoning by appealing to epistemic norms, and responsible belief-forming processes. We do have a sufficiently firm grasp of evidential support independently of explicitly causal concepts. In addition, there are accounts of objective chance that deliver evidential relations without appealing to causation—Lewis' Best Systems Account (1994) and related statistical-mechanical accounts (Loewer 2001).

Finally, a fourth concern with circularity relates to the move from evidential reasoning in general to the kind of evidential reasoning of use to deliberating agents. The concern is that we must appeal to causes to explain why certain evidential relations are undermined by deliberation—particularly evidential relations towards the past. For example, we might argue that deliberation undermines certain correlations because it must satisfy Woodward's or Pearl's *causal conditions* on being an intervention. Or we might argue that we can't decide on past states of affairs, because the causes of decisions (our beliefs and desires) interrupt the causal correlations that otherwise obtain between our actions and past states of affairs (a view I consider in chapter 5, section 5.3d). In chapter 5, I'll avoid this concern by explaining undermining in non-causal terms.

Overall, circularity isn't a concern for the deliberative account offering an interesting conceptual elucidation, or a justification of accounts. We have a sufficiently independent grasp of the relevant evidential and deliberative notions, such that the evidential biconditional remains substantive and illuminating. These concerns do put pressure on response-dependent accounts of causation that take deliberation to be more fundamental than causation. But I haven't argued for such an account. My project has been to relate causation, deliberative and evidence in a way that explains why causal relations matter.

## Conclusion

In this chapter, I've argued that prominent accounts do not adequately fulfil key desiderata on accounts of causation. While interventionist accounts from Woodward and Pearl attempt to connect causation to our needs as agents, they do so only in a trivial manner. And these accounts don't relate causation to fundamental laws. Lewis' reductivist account does better at relating causation to laws, but at the cost of using standards that appear *ad hoc* and unmotivated. We should instead adopt a deliberative account that takes causal relations to correspond to the evidential relations we use when we decide on one thing in order to achieve another. This deliberative account gives necessary and sufficient conditions on causation, elucidates a developed concept of causation, constrains metaphysical accounts, explains why causation should matter to us, and helps relate causation to fundamental laws. And it does so while avoiding circularity concerns, and delivering causal relations that are suitably objective. In the next chapter, I go on to consider how the account can also explain the temporal asymmetries of causation and deliberation.

# Chapter 5

## Temporal Asymmetries

*…no one chooses to have sacked Troy; for no one deliberates about the past, but about what is future and capable of being otherwise.*

Aristotle, *Nicomachean Ethics* (VI, 2)

When we deliberate about what to do, we deliberate on *future* states of affairs—we think about what we're going to do later today, tomorrow, or next week. We don't deliberate on *past* states of affairs in the same way. While we might regret what we did and think about whether we *should* have done otherwise, we don't deliberate on what is to be the case yesterday or last week. We don't hold the past open to choice Practical deliberation is temporally asymmetric. Call this the deliberative asymmetry.

Causation is also temporally asymmetric. In everyday life and in science, causes come prior in time to their effects. The causes of a train collision or an unhappy meal lie in prior states of affairs, not future ones. We might not want to rule out backwards causation *a priori*, particularly at the microlevel. But for macroscopic states of affairs of the type we ordinarily observe, causation does exhibit a marked temporal asymmetry: causes come before their effects. Call this the causal asymmetry.

In this chapter I'll argue that these temporal asymmetries of causation and deliberation are related—but not in the way one might expect. Typical approaches explain the deliberative asymmetry in terms of causal asymmetry: we deliberate on the future *because* our deliberation and decisions can only affect the future. In this chapter, I take a different approach. I'll argue that causes come prior to their effects *because* our deliberations are always oriented towards the future. Because we deliberate before we decide, we can only make use of evidential relations directed to the future. So we deliberate only on the future. And given that causes pick up on the evidential connections we use in deliberation (chapter 4), causes come prior in time to their effects.

Explaining causal asymmetry in this way relies on giving a non-causal explanation of why we deliberate before we decide. The epistemic model of deliberation (chapter 2) combined with an epistemic asymmetry does this work. The epistemic asymmetry I'll appeal to is the fact that we have records of the past, particularly in the form of memories, but no records of the future. This epistemic asymmetry has itself been explained in terms of the low-entropy initial condition of the universe. My explanation of causal asymmetry therefore relies ultimately on a low-entropy boundary condition—a similar posit to the one statistical-mechanical accounts use to explain causal asymmetry. But even though the deliberative account uses resources from statistical-mechanical accounts to explain causal asymmetry, this doesn't prevent the deliberative account doing important explanatory work. In fact, I'll argue, what is right in statistical-mechanical accounts should be understood in evidential and deliberative terms.

The chapter proceeds as follows. In section 1, I argue that there's a problem explaining the temporal asymmetry of causation, which neither interventionist accounts nor Lewis' account successfully addresses. In section 2, I consider statistical-mechanical explanations of causal asymmetry. I argue that these also fail, but insofar as they are successful, their success relies on agential and evidential features. In section 3, I use the deliberative account of causation to explain causal asymmetry. Finally, in section 4, I use the epistemic model and an epistemic asymmetry to explain why we deliberate before we decide.

## 5.1 The Temporal Asymmetry of Causation

In ordinary cases we encounter, causes come prior in time to their effects. There might be possible worlds where causes follow their effects, such as when time travellers emerge from their machines before entering them. And there might be backwards causation in the actual world on the microscale—such as in quantum mechanical phenomena—or in situations we don't ordinarily encounter—such as Newcomb's problem (considered section 5.3d). But when we restrict ourselves to ordinary macroscopic situations in the actual world, causation is temporally asymmetric.

The temporal asymmetry of causation is stronger than two related features of causation. Firstly, that causation is *not* symmetric—that **a** is the cause of **b** does not imply that **b** is the cause of **a** (unlike the sibling relation). Secondly, that causation is *asymmetric*—**a** is the cause of **b** implies **b** is *not* the cause of **a** (unlike the liking relation). The temporal asymmetry of causation requires causation to be asymmetric, and therefore not symmetric. But temporal asymmetry also requires temporal orientation: if **a** causes **b**, **a** is temporally prior to **b**. The temporal asymmetry of causation is the only asymmetry of causation I will be discussing in this chapter, so I will often refer to it simply as 'causal asymmetry'. There may be other ways of defining the temporal asymmetry of causation. I deliberately leave this definition broad in order to avoid presupposing a particular explanation of causal asymmetry.

### 5.1a Initial Attempts

Some straightforward attempts to explain causal asymmetry fail. Because these attempts have been persuasively criticised before, I discuss them only briefly—for further details see Horwich (1987), Price (1996), Field (2003) and Loewer (2012).

As a first attempt, one might explain causal asymmetry by appealing to an asymmetry in the fundamental laws—for example, that laws *produce* future states of the world out of previous states. Maudlin defends such a view, tracing this asymmetry back to an intrinsic asymmetry in time itself, of a type that 'underwrites claims about one state 'coming out of' or 'being produced from' another' (2007, p. 110). The problem with this suggestion is that we have no straightforward evidence that laws have an asymmetry of the required kind. As Maudlin is well aware, if we look to the mathematical form laws take in fundamental physics, laws are not temporally asymmetric in the way they determine events. They're typically time-reversal symmetric. Newtonian laws, for example, function equally well in *both* temporal directions; they necessitate the state of a system at another time, given the state of the system at *any* other time. As Russell notes, 'the future "determines" the past in exactly the same sense in which the past "determines" the future' (1912, p. 15). While there are versions of quantum mechanics that employ temporally asymmetric laws, it is notoriously difficult to apply causal concepts to such theories at the microscale. And there is no reason to think the asymmetry would scale up to a macroscopic causal asymmetry that is sufficiently general—quantum mechanics is required to recover much of Newtonian mechanics at ordinary length and energy scales. Similar arguments apply to other violations of temporal symmetry, such as in neutral K-meson decay (*pace* Dowe 1992b). These violations are too rare to explain the pervasive causal asymmetry we observe.

To defend the claim that the fundamental laws are temporally asymmetric, Maudlin is led to claim a metaphysical asymmetry in the laws that is not based in how these laws are formulated in physics. But even if we accept a metaphysical asymmetry, an explanation of causal asymmetry will still have to go via *physical* (not metaphysical) features of the world—and so still faces the same problem of reconciling the asymmetry of causation with the apparent physical symmetry of laws. A metaphysical asymmetry alone, unconnected to actual physical occurrences is not enough. To see why, imagine a time-reversed metaphysics, where the future 'produces' the past. If such metaphysical laws make no difference to actual physical occurrences, so that we still have what looks to be forwards causation in this world, then our original metaphysical asymmetry can't have been enough to explain causal asymmetry. Unless changing the metaphysical laws makes a difference to actual physical occurrences, metaphysical asymmetry can't explain a physical asymmetry like causation.

The temporal symmetry of the fundamental laws might instead lead one to think that causation is also temporally *symmetric*, directed equally towards the past and future. Can one then avoid having to explain causal asymmetry? Effectively, no. We would still need to explain why it *seems* causation is asymmetric and why we apply causal concepts asymmetrically in ordinary and scientific cases. So we would still face an equivalent explanatory challenge. Moreover, if we presume CAUSATION is a useful concept, whatever resources we use to explain why we employ the concept asymmetrically can equally be used to explain why causation itself is temporally asymmetric.

Another attempt takes the asymmetry of causation to be analytic, such that it is part of the concept CAUSATION that '**a** is a cause of **b**' implies '**a** comes before **b**'. Hume takes roughly this line, claiming that the idea of causation derives in part from the idea of temporal 'priority' (1739, Book I). But this simply shifts the explanatory burden. We would then need to explain why we use the concept CAUASTION (in which causes precede their effects) rather than the concept CAUASTION* (in which effects precede their causes). Taking the asymmetry of causation to be a matter of convention does not resolve the puzzle.

Perhaps we could take the asymmetry of causation to be a primitive feature—something discovered empirically but not further reducible or explainable. This is effectively the approach taken by Woodward (2003) and Pearl (2000). Woodward requires some direct causes to be taken for granted when working out other causal relations, and does not attempt to explain why causes come before their effects. He instead adopts an anti-reductionist stance that denies we should explain causal relations in general, beyond relating them to counterfactuals concerning other causal relations. By proceeding entirely at the causal level, his account does not explain the temporal asymmetry of causation.

Pearl's position is more complicated. When deriving causal models from statistical data, he thinks we can often assume causes come prior to their effects (2000, p. 56). But he also explains causal asymmetry by appealing either to a temporal asymmetry of agency (2000, p.111), or to the fact that we *take* certain variables to be exogenous (independent) and others endogenous (dependent) (2000, pp. 349–50). These are agent-based explanations. I'll consider them below. Pearl also notes that, if we characterise causation using probabilistic relations, causal direction depends on which variables we use to model a system—see also Field (2003). Pearl suggests evolution may explain why we select variables such that the causal order aligns with temporal order—*predictions* of future events being more useful than *explanations* of past events (2000, p. 59; Pearl and Verma 1991). But unless we assume primitive asymmetries in explanation and prediction, this will not help us account for causal asymmetry. We have to simply assume we can't explain the future or predict the past.

### 5.1b Lewis' Account

Reductionist accounts of causation are better placed to explain causal asymmetry. Since they reduce causation to regularities and contingent features, they can rely on asymmetries in these to explain causal asymmetry. David Lewis, in particular, attempts to derive causal asymmetry from an asymmetry of 'traces' (1979a). Although unsuccessful, his account provides a useful model for how a successful explanation might go.

Under Lewis' account, two events are causally related just in case there is a chain of counterfactual dependencies between them (see chapter 4, section 4.3a). So any causal asymmetry must rest on a counterfactual asymmetry: that the past does not depend counterfactually on the present, but the future does. Lewis attempts to derive this counterfactual asymmetry from an asymmetry of *miracles* (1979a, p. 473–4). According to Lewis, when we evaluate counterfactuals, we introduce a small violation of the actual laws (a miracle) that allows the counterfactual antecedent to be satisfied without further violations of the laws. Lewis argues that it takes vastly fewer miracles to satisfy an antecedent if we introduce a miracle *before* the antecedent, compared to introducing a miracle afterwards. If we use standards for evaluating counterfactual that require us to minimise the number of miracles, while securing the largest possible region of perfect match between worlds, we will posit miracles *before* the antecedent and not after. This means events before the miracle will remain exactly the same in counterfactual worlds, while events after may be massively different (see figure 6, p. 138). So, the future depends counterfactually on the present, while that past (before the miracle) does not.

Lewis derives this asymmetry of miracles from what he calls an asymmetry of *overdetermination*: the fact that events leave *traces* in the future, but not the past. For example, say, in the actual world, Nixon stands before a button. If he were to press this dread button, nuclear war would ensue. Thankfully, he does not press it. Nixon's not pressing the button leaves many records or 'traces' in the future: Nixon's memories and sense of relief, no signal in the wire, no buildings destroyed, no lives lost. Take it that *any* of these traces is enough, combined with the laws, to determine that Nixon does not press the button. If so, Lewis argues, introducing miracles after the antecedent would require using a large number of miracles to remove all these traces and produce a world which evolves *backwards* into one in which Nixon *does* press the button: 'how could one small, localized, simple miracle possibly do all that needs doing?…I put it to you that it can't be done!' (1979a, p. 473). Because events leave many traces in the future, but not the past, miracles are always better placed *before* the antecedent, and not after. So events before the miracle don't differ from the actual world, and the past does not depend counterfactually on the present.

But there's a problem: there is no asymmetry of overdetermination of the type Lewis posits. The following argument is inspired by Elga (2001) and Albert (1994), whose arguments I consider shortly. While events do leave traces in the future, these traces are not enough, individually or together, given the laws, to determine past events. As we saw in Russell's critique of causation (chapter 1), nothing less than information concerning every part of a system determines whether an event at another time occurs, given the laws. Unless we specify information about the location and velocity of *every* particle in the system, we cannot guarantee a world in which Nixon does or does not press the button. Simply keeping traces intact does not mean the past is kept intact—other parts of the system can always undermine any particular trace's reliability. So we don't need to introduce further miracles to obliterate all the traces of Nixon's not having pressed the button—a single miracle *after* the antecedent might well be enough to lead to the satisfaction of the antecedent. If so, Lewis' account won't explain counterfactual asymmetry.

So far my argument has been directed against Lewis' *reasons* for thinking his account will deliver counterfactual asymmetry. But we haven't yet shown that his explanation of counterfactual asymmetry actually fails. We haven't seen a case in which his algorithm for evaluating counterfactuals implies counterfactual dependence of the past on the present. And it might seem that his account could still recover a probabilistic version of counterfactual asymmetry. Perhaps traces render the occurrence of a previous event *probable*, even if they are not enough to determine it. Perhaps the past *probably* doesn't depend on the future. However, as Elga (2001) argues, following Albert (1994), Lewis' account doesn't lead to a probabilistic version of counterfactual asymmetry. And we can produce actual cases where Lewis' account does fail to deliver counterfactual asymmetry. In demonstrating how, I'll appeal to results from statistical mechanics, results that lead to a more promising explanation of causal asymmetry.

Begin by distinguishing between the macrostate of a system and its microstate. The macrostate of a system is characterised by its macroscopic properties such as temperature, pressure and volume. The microstate of a system is given by the exact configuration of its microconstituents—the locations of all the atoms, or other particles, and their momenta. Typically a macrostate is compatible with many different microstates. We can represent microstates by points in *phase space*—a continuous space containing 6 dimensions for every particle in the system (one for each direction of position and momentum). Macrostates are then represented by volumes of phase space. Employ a statistical postulate that takes probabilities of volumes of phase space, at a particular instant, to be uniform over the standard Lebesgue measure. This implies, roughly, that for any two equally sized volumes of phase space that are compatible with the macrostate a system, the system is as likely to be in one as the other. If we were working with a finite number of microstates, this would be equivalent to taking all the microstates to be equally probable. Using this statistical-mechanical framework, we can then derive the probable behaviour of an appropriately isolated system in a given macrostate. We consider the volume of microstates compatible with the macrostate, and apply the probability measure over this volume. We evolve the microstates according to the dynamical laws—to simplify, take these to be Newtonian. We then end up with a probability measure over final macrostates of the system. So we can identify the probable macro-behaviour of the system.

The advantage of this statistical-mechanical framework is that it gives us well-defined probabilities for how a macroscopically characterised system will evolve. These probabilities are also scientifically respectable, because they play an important role in statistical-mechanical explanations; they are used to explain the second law of thermodynamics—why systems at non-maximal entropy evolve to higher entropy. The Boltzmannian entropy of a system is given by the measure of microstates compatible with its macrostate. Macrostates with larger measures of compatible microstates are higher entropy, and those with a smaller measure of compatible microstates are lower entropy. It turns out that the overwhelmingly large measure of microstates compatible with a macrostate of an isolated system at non-maximal entropy evolves into states of higher entropy. And this explains *why* the entropy of isolated an isolated system at non-maximal entropy increases over time.

Given we have these well-defined statistical-mechanical probabilities, perhaps they can be used to recover a probabilistic version of Lewis' counterfactual asymmetry? Perhaps when a counterfactual antecedent is satisfied, the past will *probably* not change. Perhaps the macroscopic traces that exist in the actual world are enough to make the past events they record *probable*. If so, introducing miracles before the antecedent may still be preferable to introducing miracles afterwards, and we may have good probabilistic reasons to keep the past intact in counterfactuals.

Adam Elga (2001), following David Albert (1994), argues that even this probabilistic asymmetry of overdetermination does not hold. Consider a world that is macroscopically identical to the actual world at times *after* a small miracle, and so contains almost all the macroscopic traces of an event in the actual world. For example, take the macrostate $M_2$ at $t_2$, which contains all the traces of Nixon having pressed the dread button at $t_1$. If we evolve the vast majority of microstates compatible with $M_2$ back in time, we don't end up in a macrostate in which Nixon presses the button. We instead end up in a world in which Nixon, his aides, the nuclear machinery and the rest of the Earth at $t_2$ have evolved in an apparently miraculous and conspiratorial way from a more disordered and higher entropy state at $t_1$. The evolution of the actual world going backwards in time is *very* fragile—small changes of the type miracles introduce will mostly lead to worlds with radically different histories from our own. Traces do not even probabilistically overdetermine events in the past.

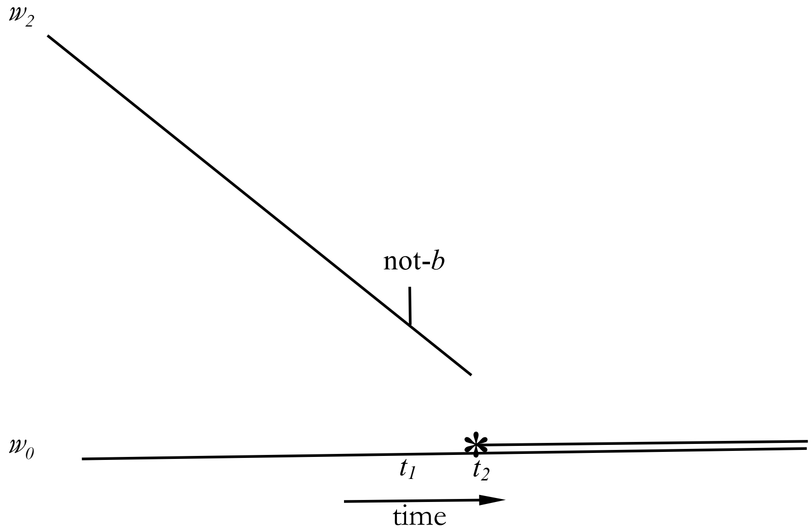


Figure 11: A counterexample to Lewis' explanation of causal asymmetry. The actual world ($w_0$) compared to a nearby possible world ($w_2$). The worlds match perfectly at all times *after* $t_2$, at which a small miracle (*) takes place, leading backwards in time to the satisfaction of the antecedent (not-*b*) at $t_1$ and massive divergences earlier.

Because of this, we can easily satisfy at least some antecedents by introducing miracles afterwards. Say, in the actual world, Nixon does press the button. Counterfactual antecedents like 'it not being the case that Nixon pressed the button' (not-*b*) can easily be obtained by introducing a miracle at $t_2$ (see figure 11). This miracle, whether it moves us to a different microstate within the same macrostate at $t_2$, or changes the macrostate in some small way, will likely be part of worlds where Nixon and his aides at $t_2$ have evolved in some conspiratorial way from a much higher entropy state. These are worlds where it is *not* the case that Nixon presses the button, satisfying the antecedent. But the state of the world at times *after* the miracle will still be exactly the same. So we have a case where introducing a miracle after the antecedent satisfies Lewis' conditions just as well as introducing a miracle before. So the past does depend counterfactually on the present, and Lewis' explanation of causal asymmetry fails.

The problem with Lewis' explanation is that Lewis only considered worlds that are macroscopically normal: worlds in which macroscopic traces of the past are reliable, and where particles don't conspire to produce traces in the future. But the dynamical laws are not enough to guarantee worlds like this. Once we learn what *does* account for the normal macroscopic behaviour of the world, we'll have the resources to consider more successful statistical-mechanical explanations of causal asymmetry.

### 5.2 Statistical-Mechanical Accounts of Causal Asymmetry

Statistical-mechanical accounts of causation are variations of Lewis'. These accounts also reduce causation to counterfactual dependence, evaluate counterfactuals using small changes from the actual world and appeal to an asymmetry of overdetermination. But they offer an account of how this asymmetry of overdetermination arises, appealing to entropy—and so are able to counter Elga's objection to Lewis. They also use more scientifically motivated methods for evaluating counterfactuals, rather than Lewis' primitive similarity relation. A guiding idea of these statistical-mechanical accounts is that casual asymmetry is to be explained in scientific terms—by looking to the methods and resources that are used to account for other physical asymmetries, such as the asymmetry of entropy. As Barry Loewer puts it, we get 'an explanation in terms of physics—a *scientific* explanation—of why we can affect the future but not the past' (2007, p. 320, his emphasis). Albert (2000, 2014, 2015), Loewer (2007, 2012) and previously Kutach (2002, 2007) all defend accounts of this form.

But despite these improvements, problems remain with these statistical-mechanical accounts. I'll argue, firstly, that these accounts don't adequately *justify* their methods of evaluating counterfactuals—and, insofar as they do, they rely problematically on notions of primitive control. Secondly, these accounts only rule out backwards causation if they introduce further epistemic and pragmatic requirements on causal relations—they don't deliver a purely scientific explanation of causal asymmetry, as they seemed to promise. Thirdly, even these additional requirements are not enough to deal with an important class of counterexamples—cases in which agents directly control records of the past. Through considering these objections, I'll argue that what is right about statistical-mechanical accounts should be understood in evidential and deliberative terms—terms that allow us to give a more straightforward account of causal asymmetry.

### 5.2a Statistical-Mechanical Counterfactuals

Statistical-mechanical accounts of causal asymmetry improve on Lewis' in significant ways. Firstly, they avoid Elga's objection. The source of Elga's objection was that when we evolve counterfactual worlds backwards in time, we end up in much higher entropy states. These are worlds that evolve conspiratorially forwards in time and so include 'fake' traces. Statistical-mechanical accounts avoid this problem by stipulating that counterfactual universes *begin* in low entropy. The entropy at the start of the universe must be low enough such that it increases all the way to the present—in accord with what we expect from the second law of thermodynamics. Call such a low-entropy initial state a 'Past State'. Statistical-mechanical accounts also constrain counterfactual worlds to begin in the *particular* low-entropy macrostate that the actual universe begun in.[104] Albert calls this constraint the 'Past Hypothesis' (2000, p. 96).

Adding in either the Past Hypothesis or a Past State as a constraint on counterfactual worlds might seem *ad hoc*. But these constraints are well motivated in science, and are used in scientific explanations of physical phenomena. Boltzmann famously appeals to similar low-entropy states to explain the second law of thermodynamics (Sklar 1993, ch. 10; Horwich 1987; Reichenbach 1956). Later followers in this Boltzmannian program appeal more directly to either a Past State or the Past Hypothesis to explain the second law (Feynman 1965; Albert 2000, 2015; Price 2002; Callender 2004; Greene 2004; Penrose 2005; Loewer 2007, 2012; Carroll 2010). The general form of this explanation is as follows.

Begin with phase space, the probability distribution over the standard measure (the 'statistical postulate') and the dynamical laws of Newtonian mechanics. As before, we can derive the fact that the entropy of an isolated system is likely to increase towards a maximum towards the future. This is because the overwhelmingly large measure of the phase space compatible with a non-maximal entropy macrostate lies on trajectories that increase in entropy towards the future. But because Newtonian laws are symmetric, this reasoning works equally well in *both* temporal directions—so it seems like we should infer that systems also begun in higher entropy states. We end up in worlds like Elga's, where particles apparently conspire to produce lower entropy states. If so, we haven't explained the second law, but have instead derived the law that the entropy of isolated systems will increase in *both* temporal directions. The solution is to stipulate that the universe begun at low entropy. This

[104] We will need to amend these claims if the concept of entropy does not apply at the very start of the universe, or if the universe has no beginning—see Albert (2000, p. 85) and Earman (2006, p. 412).

rules out these conspiratorial worlds. While the actual world will then look conspiratorial evolved backwards, this is just what the actual world is like. While objections are raised to this Boltzmannian explanation (Leeds 2003; Winsberg 2004; Earman 2006), my task is not to defend its details here. The general form of the explanation is well-accepted and scientifically plausible—philosophical accounts based on it are worth taking seriously.

Statistical-mechanical accounts improve on Lewis' in a second significant way—by refining the method for evaluating counterfactuals. I'll present Albert's method (2000, ch. 6, 2014, 2015, ch. 2), noting where Loewer's and Kutach's accounts differ. To evaluate a counterfactual, take the location of the actual world in phase space at the time of the antecedent, and consider the closest world (as measured in phase space at that time) that satisfies the following:

1. The antecedent of the counterfactual is satisfied,
2. The Past Hypothesis is true,
3. The world's macrohistory, given its macrostate and the Past Hypothesis, is assigned a reasonable probability by the statistical postulate.

Take the microstate of this world at the time of the antecedent and evolve it forwards and backwards in time using the fundamental dynamical laws. The counterfactual is true if and only if its consequent is satisfied.

Here's an example. Say, in the actual world, I'm in my apartment holding a glass of water. Consider the counterfactual: if I were to tip my hand, water would spill from the glass. To evaluate this counterfactual, we consider all the worlds in which I tip my hand (condition 1).

We reject the conspiratorial worlds which begun in high entropy states (condition 2). We also reject those that include very unlikely conspiratorial behaviour at other times, and so have unlikely macrohistories (condition 3). Most of the remaining worlds are still quite different from the actual—in some of them I'm still holding my glass, but perhaps the water has turned to ice, or I'm in prison, or on a pacific island, or all my neighbours are elephants. To minimise these apparently irrelevant changes, we pick the *nearest* world in phase space to the actual world at the present time—the world that involves the least differences in particle positions and momenta. So we pick a world where I'm still in my apartment, and the glass still contains liquid water. When we evolve this state forwards in time using the laws, gravity does its usual work and water spills out. So the counterfactual is true. Albert's initial claim is that lots of future-directed counterfactual like this come out true, but hardly any past-directed counterfactuals. More on why in a moment.

Loewer (2007, 2012) employs a similar method. But rather than picking a *single* nearby world with *statistically normal* behaviour, he takes a *probability measure* (the statistical postulate) over a *volume* of nearby worlds—worlds in which the only macroscopic change from the actual is the counterfactual antecedent, but microscopic changes are permitted.[105] We then evolve these worlds and see on what measure of them the consequent is satisfied. On Loewer's account, counterfactuals will typically have probabilistic consequents. If I were to tip the glass, it is merely very probable that water would spill out, not certain.[106]

[105] On Loewer's account, simultaneous macroscopic causation is ruled out by fiat, since the macrostate of the world in the present outside the antecedent is held fixed. I raise concerns with this restriction below.

[106] Kutach (2002) explores a similar proposal, using a probability measure over worlds where the *microstate* outside a given region R containing the antecedent and other states is held fixed (2002, 2007). But he doesn't state how R is to be specified or defend the proposal in full. Nor do the variations between Kutach's account and Loewer's affect my arguments. For this reason, I only consider Loewer's account below.

Why would we expect such results? Albert and Loewer rely crucially on the claim that the Past Hypothesis produces an asymmetry of overdetermination. Both methods restrict changes to the present (the time of the antecedent). Because of this, counterfactual worlds will include lots of traces of the actual past—in the Nixon example, the signal in the wire, memories and so forth. We saw above that these traces plus the dynamical laws were not enough to guarantee that the past occurred as it did—we were lead to infer to much higher entropy states in the past. But, Albert and Loewer claim, the traces plus the dynamical laws plus the Past Hypothesis and the statistical postulate *are* enough to guarantee past events (or at least make them very likely). By adding in the Past Hypothesis, we pick up on an actual asymmetry of overdetermination. More precisely, Albert and Loewer's accounts rely on the Past Hypothesis explaining an asymmetry of 'records'—why we have records of the past, but not the future (Albert 2000 ch. 6; 2014).[107] A record, in Albert's technical sense, is a state of a system at one time that is reliably correlated with the state of a system (itself or another) at another time, given the dynamical laws, the Past Hypothesis and the statistical postulate. A photograph, for example, might be a record of what I looked like 5 years ago. My memories of not having tipped water onto the floor might be a record of my not having done so. These records allow us to make reliable inferences about what has happened at another time, without knowing the state of the system of interest now. Without knowing what I look like now, you can still use a photograph to infer what I looked like 5 years ago. Records work in this way because the Past Hypothesis effectively provides a constraint on the initial states of recording devices. Given I know the present state of the record, when reasoning to the past, I reason to a state that lies at a time *between* two constrained states. This provides me with much more information, compared to when I reason to states further in the future. Towards

[107] For criticisms of Albert's explanation of the epistemic asymmetry, see Leeds (2003), Frisch (2007), Weslake (2014). I consider some of these criticisms below.

the future, the state I'm reasoning to does not lie between the two constrained states.

If the Past Hypothesis explains why we have records of the past and not the future, and if we use the same kind of procedure to infer using records as in evaluating counterfactuals, we will get an asymmetry of counterfactual dependence. Because local states of the present are records, they remain reliably correlated with local states of the past when we evaluate counterfactuals using the dynamical laws, the Past Hypothesis and the statistical postulate. By keeping most of the present intact, including records, we prevent much of the past from changing. But there is no similar guarantee about the future. We don't have records of the future. So future changes are not constrained in the same way. Changes needed to satisfy counterfactual antecedents may well bring about large-scale changes in the future. So we have counterfactual asymmetry.[108]

However, statistical-mechanical accounts must also adopt restrictions on the counterfactuals they consider if they are to explain counterfactual asymmetry. Firstly, they must restrict the counterfactual consequents. This is particularly clear on Albert's method. Albert's method involves picking out a single counterfactual world. Because a counterfactual world necessarily has a different microstate from the actual world in the present, it necessarily has a different past and future microhistory. So necessarily the past depends counterfactually on the present, if we allow for microscopic consequents. On Loewer's account, we get other unusual results if we allow for microscopic consequents—irrelevant macroscopic changes will make it very unlikely that the past microstate was as it actually was.

[108] Neither Albert nor Loewer commit to a particular view about how counterfactual relations relate to causal relations. But they both have something broadly Lewisian in mind.

To avoid these concerns, Albert restricts the counterfactual consequents to local macroscopic states of affairs. These are elsewhere characterised as those that concern a 'relatively small and unisolated *subsystem* of the world (Napoleon, say, or Woody Guthrie, or Greece), and which can be expressed in a relatively simple, natural, straightforward, everyday sort of language' (2000, p. 122 n. 11, his emphasis). Loewer allows for microscopic consequents, but argues that counterfactual dependence of past microstates will not amount to 'control' because we can't know about the particular micro-correlations that obtain (2007, p. 318). Macrostates are a 'natural limit on the extent of accessible information' (2007, p. 317 n. 40). So even if we don't strictly get an asymmetry of counterfactual dependence, we can still recover an asymmetry of *control*. I'll say more about the criteria for control below. Under both accounts, if the tilt of my hand is only correlated with changes in the configuration of air molecules 100 feet above my head, this does not amount to a suitable form of counterfactual dependence to endanger the explanation of causal asymmetry.

Albert and Loewer must also restrict the counterfactual antecedents. If they don't, large-scale antecedents may be counterfactually correlated with macroscopic changes in the past, and we won't get counterfactual asymmetry. For example, if the antecedent were 'New York is in the centre of Australia', the counterfactual worlds picked out would need to have very different past histories from our own, and we would have significant macroscopic dependence of the past on the present.[109] Albert restricts the antecedents using a 'fiction of agency': a 'primitive and un-argued-for and not-to-be-further-analyzed conception' of what features are to be thought of as falling under our '*direct* and *unproblematical* and *unmediated* control' (2000, p. 128,

[109] A similar concern faces Lewis' account, concerning counterfactual dependence during the 'transition period'—the time between the miracle and the antecedent. Lewis' own response is to accept such dependence, but argue it won't take any particular form (1979, p. 463). I'm not convinced by Lewis' response, but take it there are stronger reasons for rejecting his account.

his emphasis). This 'black-box' conception can be filled out in various ways—to give us direct control of our limbs, for example, or our tendons or the electrical nerve impulses in our brains. Albert claims that under any reasonable conception, our direct control will be localised to a very small area of the universe. Loewer restricts the antecedents to decisions, actions, or a combination (2007, pp. 316–7; 2012, p. 127). He takes decisions to be microscopic states in the brain that are correlated with motions of an agent's body in the future. The decisions 'open' to an agent while deliberating are those compatible with the macrostate of her brain. As a simplifying assumption, the open decisions are taken to be equally probable, relative to the macrostate of her brain and the environment 'prior to, and at the moment of, making the decision' (2007, p. 317).[110] Loewer employs a 'myth of agency', in which an agent has 'immediate control' over a range of decision: 'the supposed ability…to freely choose one among alternative decisions independently of anything else in the universe' (2012, p. 127). This is an ability agents lack if determinism is true.

Under both accounts, as explicated so far, by restricting the counterfactual antecedents to features that are under our apparent direct control and by restricting the consequents to readily characterisable macroscopic features, we get counterfactual asymmetry. The future counterfactually depends on the present, but the past does not. And if causation requires counterfactual dependence, we have an explanation for causal asymmetry.

### 5.2b Relying on Agency

To evaluate these statistical-mechanical accounts, I'll consider how they fare with respect to the three desiderata on accounts of causation introduced in chapter 4 (section 4.1). These

[110] By taking decisions to be only probabilistically relevant for *future* states of affairs, Loewer effectively builds in an asymmetry by hand. I use this as an objection to Loewer's account below, section 5.2e.

accounts do well, but, I'll argue, only by relying on agential features. Despite their scientific aspirations, these accounts do not provide a more objective alternative to agent-based approaches.[111]

Begin with the third desideratum—relating causation to fundamental laws. Here statistical-mechanical accounts improve on Lewis'. Recall that Lewis' method requires violating the actual laws (at our world) *within* the counterfactual world: a small miracle takes place. Statistical-mechanical accounts require no such violations. The dynamical laws are kept intact within counterfactual worlds. So there is an even closer relation between causation and fundamental laws.

Regarding the first desideratum, explaining why causation is useful, these accounts improve on Lewis'. Firstly, they avoid Lewis' appeal to a primitive notion of similarity with its apparently *ad hoc* standards. They appeal instead to distance in phase space, a statistical postulate and the Past Hypothesis—posits that are used in scientific explanations. So we seem to have a more scientifically respectable way of measuring departures from actuality, one that seems better placed to account for the usefulness of causation.

What about the restrictions introduced on the counterfactual antecedents and consequents—are they similarly well-motivated and objective? Statistical-mechanical accounts restrict the consequents to features that can be characterised macroscopically in relatively simple everyday language. While this restriction is somewhat vague, it is reasonable. This restriction requires us to pick out consequents of the type that feature in higher-level

[111] For other concerns with statistical-mechanical accounts of causation, see Sklar (1993), Price and Weslake (2009), Frisch (2010), Winsberg (2004) and Earman (2006).

sciences and stable macroscopic laws. If we're to explain an asymmetry in macroscopic causation, such a restriction is reasonable.

Statistical-mechanical accounts also restrict the antecedents to decisions or actions (Loewer), or to a small area over which we have 'direct control' (Albert). These restrictions are far more problematic and revisionary, and are more radical departures from Lewis' account. We effectively give up Lewis' project of accounting for a causal asymmetry in objective agent-neutral terms, since agential terms are required to restrict the antecedents. Recall, this restriction is crucial for deriving counterfactual asymmetry, because it helps keep records in the present intact. Scientifically motivated posits alone do not account for why causation is temporally asymmetric.

Could we replace the restriction with something less agential, and perhaps just restrict the antecedents to relatively small, localised areas? There are two problems with this suggestion. Firstly, we would need some motivation for it. Intuitively, causation does not just capture a relation that holds between small states of the universe, but between large and spread out states. Appealing to the limited powers of agents gave us a reason for such a restriction. Without this appeal, the restriction becomes *ad hoc*. Secondly, we would run into additional problems satisfying the second desideratum: explaining why causation is accessible to agents. Loewer's 'myth of agency' and Albert's 'fiction of agency', while somewhat vague and simplified, do at least gesture at a connection between an agent's abilities and causal relations. This feature is lost if the antecedents are restricted using non-agential terms.

What about the second desideratum: explaining how agents can, or take it they can, intervene in the relevant sense? Even with an agent-based restriction on counterfactual antecedents, these accounts don't fully satisfy this desideratum. Neither Albert nor Loewer explains how agents come by the 'fiction' or 'myth' of agency, or why such a fiction is reasonable. Albert does not even specify the scope or content of this fiction—it remains an unanalysed primitive. While Loewer suggests that free decisions should be modelled as microstates in the brain compatible with the brain's macrostate (and probabilistically independent of previous macrostates in the brain and environment), he provides no justification for this suggestion. If we are to understand how agents can intervene, we need to do better. Understanding the notion of intervention involved is particularly a concern given that statistical-mechanical accounts are counterfactual accounts: causal relations are explicated by reference to counterfactual worlds that never come to be. They face the *prima facie* challenge of accounting for why these unactualised counterfactual worlds are relevant to what we control in the actual world.

### 5.2c Relying on Evidential Features

Statistical-mechanical accounts go some way to accounting for why causation is useful and accessible to agents—but only by directly appealing to agential features. What I will argue now is that they must appeal to epistemic and evidential features as well. Firstly, they need to appeal to evidential features in order to justify why the method for evaluating counterfactual is appropriate—and so why causation is useful to agents. Secondly, they need to rely on evidential features to explain why the method delivers an asymmetry of counterfactual dependence. Altogether, *what is right in these statistical-mechanical accounts should be understood in evidential and agent-based terms.*

I argued in the last chapter (section 4.3b) that even if a counterfactual account captures the extension of our causal concept, this does not give us a philosophic justification for why we should evaluate counterfactuals in those terms. How might Loewer and Albert justify evaluating counterfactuals in statistical-mechanical terms? In particular, how might they justify holding the Past Hypothesis and dynamical laws fixed, and using the statistical postulate? Loewer claims that the Past Hypothesis and the statistical postulate are lawlike under a Lewisian conception of laws (2007, pp. 304–306). But this isn't sufficient justification. Under a Lewisian conception, laws derive from the best systematisations of the actual events that occur in our world. Given they're just ways of summarising information about actual events, it's unclear, *prima facie*, why we should expect them to be held fixed when evaluating counterfactuals about non-actual events. Lewis himself allows for violations of the actual laws in counterfactual worlds. Nor can we straightforwardly appeal to the fact that the method picks out what is 'objectively more likely' to happen (Loewer 2007, p. 323). The reasonableness of the method is precisely what's in question.

But while neither Albert nor Loewer fully justify why their methods for evaluating counterfactuals are appropriate (beyond noting their scientific credentials) they do say a few things that are suggestive. Albert claims his method of evaluating counterfactuals captures our 'normal procedures of inference' (2000, p. 129). So when we ask why we have this particular method for evaluating counterfactuals, we have a direct answer—it is because it captures the structure of how we ordinarily make inferences in the actual world. This justifies why we should keep the dynamical laws and the Past Hypothesis fixed when evaluating counterfactuals. Albert in fact goes a little further, and takes these statistical-mechanical features to be 'part and parcel of what we mean when we speak of that world as being "like ours'" (2000, p .129). While Loewer doesn't appeal to epistemic features in justifying his counterfactual method, he also takes the method to provide a scientific explanation of reliable inferences (2012, pp. 126–7). I claim it is precisely insofar as a counterfactual method picks up on good inferential reasoning that we are justified in using it in giving an account of causation.

Evidential relations also feature in explaining *how* the methods deliver a counterfactual asymmetry. Albert and Loewer claim that the Past Hypothesis, the dynamical laws and the statistical postulate explain an asymmetry of records—why we have records of the past and not the future. If we then evaluate counterfactuals using the same method we use when inferring with records, keeping the present records intact in counterfactual worlds means the past will be kept intact as well. Because the future is not constrained by present records, we then have counterfactual asymmetry. What's important to note is that the whole success of this explanation hinges on three epistemic features: there being an asymmetry of records (that remains intact in counterfactual worlds), records being kept mostly intact in counterfactual worlds, and the method for evaluating counterfactual having the same structure as inferring using records. Provided we keep most of the records intact, and we 'infer' in counterfactual worlds according to the same method we ordinarily use, an asymmetry of records straightforwardly delivers an asymmetry of counterfactual dependence. The success of the statistical-mechanical explanation depends on it being an instance of this general pattern—and it is an explanatory pattern we can see without even mentioning statistical-mechanical features such as the Past Hypothesis.

The key claim I want to make is that statistical-mechanical explanations of counterfactual asymmetry *should be understood at this more general epistemic level.* It is because of an epistemic asymmetry, and because the method for evaluating counterfactuals follows the same structure as our ordinary inferential reasoning, that statistical-mechanical accounts deliver counterfactual asymmetry. This gives us the right level of explanation, and the right way to understand the success of statistical-mechanical explanations (insofar as they are successful).

Moreover, keeping things at this general level allows us to bracket some concerns about the statistical-mechanical explanation of the epistemic asymmetry. Frisch (2007) and Weslake (2014) concede that the Past Hypothesis (or simply a low-entropy Past State) plays an important role in explaining why there is epistemic asymmetry, and is perhaps necessary for records. But they doubt it is sufficient.[112] If the Past Hypothesis is necessary but not sufficient for records of the past, and there is no Future Hypothesis, we still have a partial explanation for why we have records of the past and not the future—past records are possible, future records are not. But what we don't have are sufficient features to guarantee the past is kept intact when evaluating counterfactuals using statistical-mechanical methods. And these missing features would need to be included if the methods were to deliver counterfactual asymmetry.

This concern can be pressed as follows. Albert is explicit that the Past Hypothesis involves not merely a low-entropy state, but the particular past macrostate of the early universe (2000, p. 96). What explains why our inferences towards the past are explained or justified isn't simply that the universe's entropy has been increasing. Nor is it the fact that isolated

[112] Leeds (2003) and Earman (2006) doubt both its necessity and sufficiency. I leave aside concerns about how we come to *know* the Past Hypothesis. Albert's explanation may be offered as an externalist justification of what makes our inferences reliable, without us being able to infer by those same means (*pace* Frisch 2007).

subsystems of the universe, including recording devices, increase in entropy. Albert rejects Reichenbach's explanation of the epistemic asymmetry (2000, pp. 123–4). Instead, Albert explains the asymmetry of records as follows (2000 ch. 6; 2015, ch. 2). All we could possibly expect to know about the future *could* be inferred by appealing to the dynamical laws and our knowledge of the past and present (2000, p. 114).[113] Yet we take ourselves to know vastly more about the past than could be inferred from the laws and what we know of the present and the future (2000, p. 122). 'Retrodiction' of this form would give us wildly inaccurate results, leading us to infer to conspiratorial worlds that begin in higher entropy states. The Past Hypothesis rules out such conspiratorial worlds. It also allows us to reason from two macroscopically characterised states at different times to a time in between. If inferences of this form are more powerful, the Past Hypothesis seems well placed to explain and justify how we reason to the past. As Albert puts it, we need something in addition to the laws to underwrite our reliable reasoning towards the past, something that 'must be prior in time to everything of which we can potentially ever have a record, which is to say that it can be nothing other than the initial macrocondition of the universe as a whole' (2000, p. 118).

But Albert's argument isn't enough to show that the Past Hypothesis is *sufficient* (or even necessary) to explain reliable reasoning towards the past. All it shows is that something more is required than the dynamical laws and statistical postulate. To say something more is needed is not to say that the initial macrocondition of the universe is what's needed. The Past Hypothesis may be one way of ruling out the conspiratorial cases—but there might be other cases that need ruling out, meaning the Past Hypothesis is not sufficient (Frisch 2007,

[113] More precisely, 'prediction' derives the future state of a system from the dynamical equations of motion and the statistical postulate applied to the 'world's present directly surveyable condition' (2000, pp. 96, 114). This directly surveyable condition includes the current macrostate of the world, plus any micro-features of the brain of the observer that are 'directly accessible' through introspection.

Weslake 2014). And we might not need the whole Past Hypothesis at all—more local entropic restrictions of the type suggested by Reichenbach may do the work, meaning the Past Hypothesis is not necessary. While some restriction on past macrostates is required, it's yet to be shown that the Past Hypothesis is what's required. And without showing this, we don't know whether these statistical-mechanical counterfactuals are enough, or what's required, to deliver reliable inferential reasoning.

I don't take myself to have shown that the Past Hypothesis cannot account for how we infer using records. And I will later appeal to a low-entropy condition to explain an asymmetry of records (section 5.4c). But we are right to be concerned that the structure of our reasoning hasn't been explained. And until it is, we can't simply take on the statistical-mechanical explanation of epistemic asymmetry and adopt it as a method for evaluating counterfactuals. We lack adequate justification for thinking the method will lead to the right results. Instead, we should think of these statistical-mechanical accounts as *instantiations* of the deliberative account. As far as statistical-mechanical accounts are successful, their success relies on their tracking evidential relations, those, I will argue, of use to deliberating agents. They do not provide competing accounts of causal asymmetry.

### 5.2d From Counterfactual Dependence to Control

Statistical-mechanical accounts of causation rely on agential and evidential features—they restrict the antecedents using explicitly agential terms, and rely on evidential features to explain counterfactual asymmetry. In this section, I show that even these features are not enough to rule out all forms of backwards counterfactual dependence. Albert and Loewer end up also restricting the counterfactuals to those agents can know about and use for gain, appealing to a pragmatic notion of 'control'. Only by relying in these further ways on agential and epistemic features can these accounts hope to deliver causal asymmetry. For these additional reasons, statistical-mechanical accounts are not more objective alternatives to agent-based accounts.

Recall, statistical-mechanical accounts rely on keeping records of the past intact when evaluating counterfactuals. It is because we keep the present intact, and the present contains records of the past, that the past is kept intact in counterfactuals. But what about when there aren't records of past events? Elga (2001) and Kutach (2002) raise the following counterexample.[114] Say there is a large-scale event in the past, the sinking of Atlantis, for which there are no records in the present. This means that small changes in the present, such as the movement of my finger, may well be correlated with whether Atlantis exists, under statistical-mechanical accounts. Without records, the past is extremely fragile with respect to small changes in the present. Given we control small changes in the present, statistical-mechanical accounts imply we control past macroscopic events.

Albert (2014; 2015 ch. 2) and Loewer (2012, p. 128) respond to this case by introducing further conditions on what counts as the right kind of counterfactual dependence. They argue that not all forms of counterfactual dependence constitute *control*. Control requires counterfactual dependence we can know about. We can't know about the microscopic correlations between the movement of my finger now and the sinking of Atlantis. And there are no macroscopic correlations to be known—if there were, there would be records in the present. So we can't know about the counterfactual dependence of the past on the present,

[114] The case is based on ideas in Elga (2001), and is cited as Elga's in Loewer (2007, pp. 318–9; 2012, p. 128).

and we don't control the past.

Loewer also offers an additional response to this case. Because his method requires decisions to be probabilistically independent of previous macroscopic states, the probability of Atlantis' sinking cannot be changed by a decision in the present. Decisions can 'neither destroy nor create records' (2007, p. 318), and so cannot alter the probability of events like Atlantis' sinking. But this response relies crucially on an asymmetric characterisation of decision: that decisions are probabilistically independent of past, but not future, macrostates. If causal asymmetry is to be derived rather than presupposed, Albert and Loewer's above response is to be preferred. We need to explicitly include a knowledge condition on control.

Frisch (2010) also raises a counterexample. He considers a case where the unique record of an event in the present is under an agent's direct control. Say I am playing a piano piece, and come to a point in the music where I can decide whether to play the first or second ending—the decision being under my direct control. Stipulate that I have no conscious memory of what music I have played and there are no external records present. But I am a reliable pianist. Whatever decision I make *counts* as a record of what music I have played. So by having control of my decision in the present and my decision being a unique record, I have control over what music I have played in the past. So even with a restriction to control, statistical-mechanical accounts do not explain counterfactual asymmetry.

In earlier work, Albert dismisses such cases. He claims records of the past are 'manifestly *not* things that can by any stretch of the imagination be thought of as among the features of the present condition of the world that fall under our *direct* and *unmediated* and *unproblematical* *control*' (2000, p. 130, his emphasis). But in cases where decisions are reliable responses to the past, they do count as records. Loewer's 2007 account explicitly rules out such counterexamples. But it does so by ruling out correlations between decisions and past macrostates *a priori*. This is an asymmetric assumption we are not entitled to if we're to explain causal asymmetry. Albert (2014) later responds to Frisch's case by adding another condition on control. He argues that even if the agent knows she can influence the past, she cannot use this influence to gain any further benefit, beyond what she secures through influencing the present. If I want to finish playing sooner, for example, I could influence the past so that I began the piece earlier. But I could just as easily decide to play the second ending and finish playing. The past event only influences the future via a feature I already have direct control over. So Frisch's is not a case of *effective* control.

But this response fails. Albert only considers additional benefits in the future, and this introduces an unexplained asymmetry. If we consider benefits in the past as well, influencing the past in the Frisch case may give me additional rewards. For example, if I started playing earlier, I may have spent less time bored at the piano. Influencing the past is essential for securing this gain. For Albert's response to work, this asymmetry in what benefits we consider must be explained in a non-circular fashion. And even if this can be done, I'll give a further case below against which this response is ineffective.[115]

A better response is also available. To effectively control the past, it is not enough to know that in some cases, or in general, my decision is correlated with past events. To effectively

[115] Albert later revises his response, arguing that the control of the past we have in the Frisch case is not suitably robust (2015, ch. 2). In cases where we attempt to exploit the correlation to gain a reward, the correlation is screened off by the decision to gain the reward. This response agrees with my own (forthcoming$_1$). In this chapter, I go more deeply into why this screening off occurs.

control the past, I must have some way of checking whether I have controlled it, in relatively similar cases when I *attempt* to do so. I cannot do this in the Frisch case. There is no time at which I know both my decision and my past playing by independent means. Because my decision in the present is the only record of the event, I cannot determine whether I have actually controlled the past in this or similar cases. And so this is not control I can know about. Overall, statistical-mechanical accounts can respond to counterexamples like Frisch's and Elga's—but only by introducing further epistemic conditions on what counts as a suitable form of counterfactual dependence—*effective control.*

### 5.2e Time, Flies and Control

The counterexamples above can be dealt with. If we restrict ourselves to counterfactual dependence we can know of and use, it seems statistical-mechanical accounts may explain counterfactual asymmetry. But there is a third counterexample that can't be dealt with by the conditions introduced so far. I provide further details and defence of this counterexample elsewhere (forthcoming$_1$), where I use it as an argument against Albert's explanation. Here I only briefly present the case, because I am less interested in whether one can respond to it, and more interested in the form that these responses must take. The case must be responded to in epistemic and agential terms.

The counterexample is as follows. In the actual world, a fly flies in front of my face at $t_1$, and two seconds later at $t_2$ I swat it away. Then consider the counterfactual: if I had not swatted at $t_2$, the fly would have been somewhere else at $t_1$. If this counterfactual is true, and I can knowingly exploit this counterfactual dependence, I can control the past. Stipulate that I'm an exceedingly reliable fly-swatter, well-trained to swat away flies whenever they appear. This reliability may even be well-recorded. If so, my decision to swat is a *record* of the fly's location. Take it there are no other records of the fly's location in the present, except the fly itself and a tiny recording device that tracks its location. Under statistical-mechanical accounts, it seems I can control the fly's location. I directly control my decision, and because my decision is reliably correlated with the fly's location, I control the fly. Because the other records of the fly's location are small, they can be easily changed in counterfactual worlds, so as to keep the various recording devices accurate and the evolution of the world sufficiently probable.

What produces the counterfactual dependence of the past on my decision is the fact that my decision is a *record* of the past. But the crucial difference between this case and Frisch's is that in this case there are other records as well of the past as well. And it is these other records that allow me to know about my control of the past. In this fly case, the recording device can be used to check whether I do successfully control the location of the fly on occasions when I attempt to do so. (The device even allows me to be rewarded further in the future when I do so). So none of the above responses to counterexamples will work here. We seem to have a case where I can effectively control the past.

One might respond to the fly case by adopting a different method for evaluating counterfactuals. Loewer, for example, avoids the fly case by stipulating that there cannot be correlations between an agent's present decisions and macrostates in the past or present environment (2007, p. 317). This means that records like the device or the fly cannot depend counterfactually on an agent's decisions. But Loewer's stipulation is problematic for two reasons. Firstly, it rules out simultaneous counterfactual dependence *a priori*. Insofar as the temporal direction of causation is taken to be contingent and deserving of explanation, simultaneous causation should be similarly contingent. Price and Weslake (2009, p. 426) and Frisch (2010, p. 29) raise similar concerns. Secondly, an agent's decisions *are* often correlated with past and present macroscopic states in the environment—such as when my decision to buy a chocolate ice cream is correlated with the presence of an ice cream van. Someone might infer to the presence of the van by learning of my decision. If the counterfactual method is justified because it picks up on good inferential reasoning, it should allow for such correlations. And if we do rule out such correlations, we should justify why this restriction is appropriate.

My suggested response is to instead appeal to the same kind of epistemic and agential features that were used in response to other counterexamples. Responding in epistemic terms means we can keep the inferential justification for why statistical-mechanical methods are appropriate for evaluating counterfactuals. But it means we will also have a response to the counterexample that is available to defend a deliberative account. I'll give my preferred response to the fly case in the next section.

Having considered statistical-mechanical counterfactual accounts, and the counterexamples they face, we've seen that these accounts rely crucially on epistemic and agential features throughout. These accounts begin with primitive notions of direct control or myths of agency. They appeal to epistemic relations to justify why the method of evaluating counterfactual is appropriate and to explain counterfactual asymmetry. And they rely on further epistemic and agential requirements in order to distinguish 'effective control' from mere counterfactual dependence. By the end, we have moved quite far from Lewis' aspiration to explain counterfactual and causal asymmetry in objective terms that make no reference to agents. Although statistical-mechanical accounts begin in science, and secure important connections between temporal symmetries, fundamental laws and causation, they must appeal to evidential and agential features. So it seems we may be able to explain causal asymmetry entirely in evidential and agential terms. It is to such an account I now turn.

### 5.3 An Evidential Explanation of Causal Asymmetry

According to the deliberative account of causation, causal relations correspond to evidential relations. Initially it might seem that such an account won't be able to explain causal asymmetry. Evidence, as I've defined it, generally licenses inferences equally well towards the past and future. Rain in the afternoon is evidence of clouds in the morning as well as puddles in the evening. So it seems the account will lead to causation that runs both directions as well. In this section, I argue that the deliberative account *can* explain causal asymmetry. I'll argue that decisions aren't evidence for past states of affairs when we decide for the sake of the past in *proper* deliberation. If causal relations are the evidential relations we use in proper deliberation, there won't be causal relations towards the past. So we can explain causal asymmetry in evidential and agential terms.

#### 5.3a In Which Evidential Relations towards the Past are Undermined

Causation is temporally asymmetric if and only if A being a cause of B implies A is prior to B. This latter is logically equivalent to A's *not* being prior to B implying A is *not* a cause of B. To show that causation is asymmetric under the deliberative account, we need to show that this condition is met, given the evidential biconditional defended in chapter 4:

*Evidential Biconditional:* A is a cause of B *if and only if* an agent deciding A (in proper deliberation) for the sake of B would be good evidence for B.

According to the evidential biconditional, A is *not* a cause of B if and only if an agent's deciding A for the sake of B would *not* be good evidence for B. So, substituting in the above, causation is asymmetric if and only if A's *not* being prior to B implies an agent's deciding A for the sake of B would *not* be good evidence for B. In other words, causal asymmetry is explained if the following conditional can be shown to hold:

*Causal Asymmetry Conditional*: if A is *not* prior to B, then an agent's deciding A for the sake of B (in proper deliberation) would *not* be good evidence for B.

My strategy will be first to argue this conditional is plausible. In ordinary cases, we don't take the relevant decisions on A to be evidence for B. Following this, I will explain *why* this conditional holds. To begin, consider the following case.

Every morning at 8:30 am, Tamsin stops by her umbrella stand on her way out. Some days she takes her umbrella. Some days she leaves without it. On days when she takes her umbrella, there tend to be storm clouds in the sky earlier, at 8:00 am. On days when she doesn't take her umbrella, there tends not to be. Her flatmate Tod has observed her behaviour, and infers the following: Tamsin deciding to take her umbrella is good evidence of earlier storm clouds, and her *not* taking her umbrella is good evidence of their absence. And so on days when Tod is in a hurry, and has no time to look at the weather himself, he uses Tamsin's behaviour to determine whether there were storm clouds. However, this morning he finds Tamsin hesitating before the umbrella stand. Tod's in a hurry and impatient to know what decision Tamsin will make. The following exchange occurs.

> Tod: Well, were there storm clouds or not?
>
> Tamsin: I'm just deciding whether to take my umbrella.
>
> Tod: I know, but your deciding to take your umbrella is good evidence of whether there were storm clouds.
>
> Tamsin: Oh, that's easy then. I'll not take my umbrella. I don't much feel like rain.

What's gone wrong? Tamsin has chosen poorly. While Tamsin's not taking her umbrella is ordinarily good evidence that there were no storm clouds, her decision not to take the umbrella on this occasions is *not* good evidence there weren't. Somehow the correlation between her decision and the prior state has been 'undermined'. Note that this undermining does not occur merely for Tamsin, or from Tamsin's point of view. Tod should also ignore Tamsin's behaviour on this occasion, if he's interested in knowing whether there were clouds—Tamsin's decision is not good evidence of the weather, even for Tod.

Why might this undermining occur? Plausibly Tamsin's decision to take her umbrella is usually correlated with earlier storm clouds because she *observes* the weather—either directly, by looking, or indirectly, such as by hearing a weather report. Say Tamsin has observed the weather. Then her evidence already settles whether or not there were clouds. According to the conditions I've argued for, Tamsin then can't properly deliberate on taking her umbrella *for the sake of* there not being clouds. Proper deliberation requires her evidence to leave open what decision she makes, and whether the state she's deciding for the sake of obtains (section 4.4a). Alternatively, say Tamsin hasn't observed the weather. Then she can properly deliberate. But in this case, we have no reason to expect a correlation between Tamsin's decision and the clouds. The reliability of the evidential connection between her decision and the clouds depends on her having evidence of the clouds. Without this evidence, Tamsin would be merely lucky if her decision was correlated with there having been clouds.

However we specify the case, Tamsin's deciding to take her umbrella in proper deliberation isn't evidence of clouds. Either her deliberation is not proper or her decision is not evidence. So the evidential biconditional doesn't imply a causal relation between her taking her umbrella and previous storm clouds. Tamsin's case exemplifies the kind of undermining implied by proper deliberation that is needed to explain causal asymmetry. If we can show that evidential relations between an agent's decisions in proper deliberation and states of affairs in the past or present are always undermined in this way, we can explain causal asymmetry.[116] And we can also explain why agents deliberate on the future and not the past. If decisions must be good evidence of the states they are decisions for, as I have argued, this undermining prevents agents properly deliberating on the past.

### 5.3b Why Evidential Relations are Undermined

I will now argue that Tamsin's case generalises: if A comes after or is simultaneous with B, an agent's deciding A for the sake of B (in proper deliberation) would *not* be good evidence for B. To generalise Tamsin's case, I'll consider what might account for this undermining.

[116] This explanation will also rule out spurious correlations as causal, but these are not my main concern.

Could evidential undermining be accepted as a brute fact? This is unsatisfactory for two reasons. Firstly, a brute fact view gives us no guide to when we should expect undermining to occur. Secondly, a brute fact view would not enough to explain causal asymmetry. It needs to be the case that undermining occurs in counterfactual cases as well—if A comes after B, an agent's deciding A for the sake of B *would not be* good evidence for B. Unless we know what this undermining depends on, we can't determine whether it would also occur in counterfactual cases.

Here's how to explain undermining. Tamsin's case involves an evidential correlation that goes *via* her observing the clouds. We only expect Tamsin's decision to be correlated with the weather when this correlation is mediated by her observation, and when the evidence from this observation constitutes part of her deliberation. I claim that this mediation is characteristic of evidential correlations between decisions and states of affairs in the past—they are always mediated by other states of affairs that form part of deliberation.[117] This gives us a principled reason for thinking Tamsin's case generalises. If correlations between decisions and the pasts are always mediated by deliberative states, then if such states are absent, we should not expect the correlation to hold. If such states are present, they make deliberation on the past improper—and so no causal relations are implied.

If Tamsin's case generalises, the evidential biconditional rules out backwards causation. An agent's deliberation is proper only if her evidence leaves open whether the state she's deciding for the sake of occurs. If evidential relations towards the past are always mediated by observations or other evidence of the past, these make deliberation on the past improper.

[117] I late discuss exceptions to this, such as in traditional Newcomb cases.

When an agent's deliberation is proper, this must be because she has no mediating evidence. But in this case there is no reason to suspect an evidential correlation between her deciding and the past state. This is the correlation we need for causation. So backwards causation is ruled out.

Note, similar reasoning motivates the principle I introduce in chapter 4 (section 4.4c) for evaluating what evidence an agent would have, were she to properly deliberate. If correlations between decisions and past states are always mediated by evidence, this is something structural about our deliberation that we should hold fixed in counterfactual cases. And so it motivates a general principle for keeping such evidence fixed when evaluating counterfactuals.

### 5.3c An Appeal to Evidential Decision Theory

If correlations between decisions and past states are always mediated by observations or other evidence, the deliberative account can explain causal asymmetry. But there are problem cases. The fly case and the Frisch case above might seem to involve instinctual actions, performed in response to unconscious cues. Perhaps while they are deliberating these agents don't have evidence of the past states of affairs they are responding to. Other problem cases arise in debates in decision theory between causal and evidential decision theorists. These theorists disagree on whether an agent should choose the option that is *evidence* for a favoured outcome (evidential decision theory) or *causes* a favoured outcome (causal decision theory). Causal decision theorists attempt to show that evidential relations alone don't settle what decision an agent should make. While proponents of either view don't typically defend an agent-based account of causation, it seems that if evidential decision theory is in trouble, so is the deliberative account. Here is a classic problem case raised against evidential decision theory.[118]

Gina is deliberating about whether to smoke. She knows there is an evidential correlation between smoking and cancer—statistics show that the incidence of cancer among those who smoke is much higher than in the general population. But she knows this is not because smoking causes cancer, but because a particular gene is a common cause of both smoking and cancer, as shown in figure 12. Gina would like to try smoking, but would like to avoid getting cancer much more. She is not sure whether she has the gene that causes cancer—she has taken no tests, and has no independent evidence.

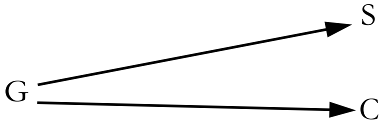


Figure 12: Causal graph showing a gene (G) as a common cause of smoking (S) and cancer (C).

Causal decision theorists argue that evidential decision theory gives the wrong advice to Gina. It seems that Gina's not smoking is excellent evidence of her not having the gene, both in terms of the objective evidential relations in the system, as well as in terms of the subjective evidential relations (Gina's beliefs) that evidential decision theory appeals to. If Gina's not smoking is (subjectively) good evidence of her not having the gene, evidential decision theory will recommend she doesn't smoke—this is the action that is evidence of a favourable outcome (not having the gene and so not getting cancer, rather than getting

[118] In this section I discuss only realistic Newcomb cases, known as 'medical Newcomb cases'—see Nozick (1969, pp. 125–30), Skyrms (1980, pp. 128ff.), Lewis (1981, p. 9ff.), Eells (1981, p. 298ff.) and Jeffrey (1981). I briefly consider how these differ from traditional Newcomb cases below.

cancer). But, it seems it would be irrational for Gina to avoid smoking for this reason. She either has the gene or she doesn't. She knows that nothing about her smoking now can affect that. So she should smoke if she prefers to. Even if her smoking is subjective (or objective) evidence of having the gene, she should not decide based on this correlation.

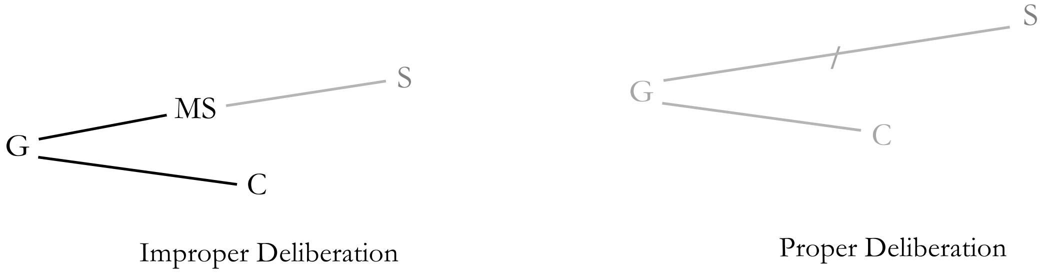
Gina's case also creates trouble for the deliberative account of causation. Gina doesn't seem to have objective evidence that already settles whether she has the gene. The correlation between the gene and her smoking is not brought about because Gina *observes* whether she has the gene by some independent means and then decides to smoke. So it seems Gina can properly deliberate on smoking even while there *is* a correlation between the gene and smoking. So the deliberative account of causation will imply that her not smoking is a *cause* of her not having the gene—contrary to the setup of the case.

In responding to such cases, evidential decision theorists have effectively attempted to show that evidential relations between past states and acts are *always* mediated by other states that form part of deliberation. The defence they give is known as the 'tickle defence', a term from Lewis (1981).[119] For example, say the correlation between having the gene and smoking is mediated by a desire to smoke. Gina's belief that she has this desire and her belief in the desire's correlation with the gene together give her (subjective) evidence that she has the gene. This evidence is independent of what decision she makes, and so it 'screens off' the evidential relevance of her act for the past state. No matter what she decides at this point, she shouldn't take her smoking to be correlated with the gene. So she may as well smoke if

[119] This response is considered in Nozick's original presentation (1969), and appealed to by Eells (1981; 1982; 1984), Jeffrey (1981), Horgan ([1981]), Price (1986; 1991; 1993) and Horwich (1987, ch. 11). Concerns with this response are raised by Skyrms (1980), Lewis (1981), Jackson and Pargetter (1985), Sobel (1994), Papineau (2001) and Joyce (2007). I consider some of these concerns below.

she wants to. If an agent always has the relevant beliefs concerning such mediating states, the evidential relevance of her decision for past states will always be screened off—and what act she decides on is no longer evidence for the past state. And so if agents decide on acts that are subjective evidence for favourable outcomes, they will not be led astray.

This tickle defence helps generalise the response I gave to Tamsin's case above and show how the evidential biconditional does not imply backwards causation. An agent's deliberation is proper only when she lacks (objective) evidence of the state she's deciding for the sake of. If evidential relations towards the past are mediated by beliefs, desires, or other states, these mediating states provide evidence of the past state and make deliberation on the past improper. When an agent's deliberation is proper, this must be because she has no mediating evidence. But in this case there is no reason to suspect an (objective) evidential correlation between her decision (or act) and the past state (see figure 13). So backwards causation is not implied.



Improper Deliberation

Proper Deliberation

Figure 13: The evidence Gina has while properly or improperly deliberating on smoking (S) for the sake of not having the gene (G), given the causal structure in figure 12. Gina's deliberation is improper if she has evidence of a mediating state (MS) that settles whether she has the gene and will get cancer (C)—even if it does not settle her smoking. Her deliberation is proper only if she lacks evidence of any such mediating state. But then her decision won't be evidence of the gene.

This response is also effective against the fly and Frisch cases. Either the agents have evidence that settles how they decide, or there is no reason to expect a correlation.

By appealing to evidential decision theory to defend an agent-based account of causation, I follow Price's lead (1991, 1993, 2012). But there are important differences between our approaches. Firstly, Price's account of causation is concerned with the credences an agent must have to correctly apply the concept CAUSATION. For this reason, he envisages a very close connection between his account of causation and evidential decision theory—both deal in subjective probabilities and evidential relations. My deliberative account appeals to more objective relations. In the next section, I argue that this is an advantage for the deliberative account. Secondly, Price uses a version of the tickle defence to defend evidential decision theory. For reasons outlined in the next section, I do not. Thirdly, Price (2012) explains why evidential correlations are undermined by appealing either to causal notions (1986) or to primitive asymmetric abilities on the part of the agent. In section 5.4, I explain undermining without appealing to these.

### 5.3d Problems for Evidential Undermining

A number of objections have been raised against the tickle defence of evidential decision theory. I consider six. Rather than respond in full, my strategy will be to show that what's needed to defend the deliberative account of causation is significantly *less* than what's needed to defend evidential decision theory. As far as evidential decision theory remains plausible, so does the deliberative account.

Firstly, what if not all correlations between acts and past states go via deliberative states? What if Gina has an addictive tendency, and smokes regardless of her beliefs and desires? This may be, but this is not the evidential relation the deliberative account appeals to. For a relation between A and B to be causal, it must be the agent's *deciding* on A that remains correlated with B, not A itself. We needn't join Eells (1981) in assuming common causes necessarily change *actions* by changing beliefs and desires.[120] Similarly, we needn't assume that all decisions are of the form 'A for the sake of B'. But it is only when decisions are of this form that the deliberative account implies causal relations.

Secondly, what if the mediating state only provides some evidence of the past, because there is only a partial correlation between the act and the past state? In this case, deliberation is still proper. But still the mediating state provides as much evidence for the past as the act does—so the evidential relevance of the act or decision for the past is still screened off.[121]

Thirdly, what if the correlation between the decision and past state holds, regardless of whether any particular mediating state occurs during deliberation? If so, there will be no screening off and the past state will continue to be correlated with the action and decision, implying backwards causation. This is what occurs in traditional Newcomb cases, where an infallible predictor has arranged states in the past to match what decisions he predicted the agent will make (Nozick 1969). In this case, the correlation between the decision and the past state survives whatever particular beliefs or desires the decision maker has about

[120] This objection is not obviously fatal to evidential decision theory either, if decision theory concerns what *deliberating* agents should do. If Gina is deliberating, she should already know she is not in the class of agents who smoke in some non-deliberative way (Horwich 1987, p. 183).

[121] Note that this response is only needed if the deliberative account of causation is extended to cover probabilistic causation—see section 4.4a (p. 155 n. 99). Otherwise the fact that the decision isn't good evidence for the past state is enough to rule out the correlation as causal.

himself, however complicated his deliberation. But ruling out backwards causation in these cases is not required to defend the deliberative account. The defence only needs to show that causation does not run backwards in realistic cases—so called 'medical' Newcomb cases. Recall, the causal asymmetry to be explained concerned cases we actually encounter. We needn't rule out backwards causation in unrealistic cases, where the epistemic capabilities of the predictor go far beyond what we ordinarily encounter. While normative evidential decision theory may need to issue a recommendation in such cases, a deliberative account does not.[122]

Fourthly, Skyrms (1980, p. 131) and Lewis (1981, pp. 10–11) have argued that the screening off defence requires agents to have too much self-knowledge. Actual agents may not have credences in the relevant beliefs and desires, even if the beliefs and desires are there. Or she may not recognise the significance of her beliefs and desires. Gina may recognise the tickle in her throat, but not recognise it as a desire to smoke. So it will not screen off the evidential correlation. A fifth concern is similar; we may want decision theory to issue recommendations not just relative to an agent's own beliefs, but relative to another agent's beliefs (Jackson and Pargetter, 1983). The tickle defence only explains why the correlation is undermined for Gina—not for an observer, who lacks Gina's evidence of her deliberative states.

[122] But a plausible response, following Price (2012), is to argue that traditional Newcomb cases do involve backwards causation (see also Nozick 1969, p. 134). The abnormal evidential structure implies abnormal causal structure. This position might be described as being a 'no-boxer': denying that we can have traditional Newcomb cases which involve no backwards causation.

These are significant problems for evidential decision theory—but not for defending a deliberative account of causation.[123] Evidential decision theory deals in subjective probabilities. A subjective notion of evidence makes sense here, since decision theory is required to issue recommendations to agents or at least explain their behaviour. But the deliberative account of causation appeals to a more objective notion of evidence. Using more objective evidential relations, puddles are evidence of rain, for example, regardless of what a given agent believes (section 4.4c). These evidential relations may be fixed by the standards of an epistemic community, or by physical chances. And they are a significant and unavoidable commitment of the kind of deliberative account of causation I am defending. For those who prefer sparser ontologies, they may be a significant cost. But these evidential relations do important theoretical work in explaining why the causal relations of the deliberative account are suitably objective. And they also help explain why ordinary cases of deliberation don't imply backwards causation.

With these more objective evidential relations, if Tamsin has seen the puddles, then she does have evidence of rain—even if she has mistaken beliefs about the correlation between puddles and rain. Similarly when Gina is deliberating, she need not recognise the tickle as evidence of the gene in order for it to provide such evidence. Agents may even have evidence when they don't believe they do. Provided Gina's tickles, beliefs and desires are directly accessible to her, they reasonable count as part of her current stock of evidence and provide her with evidence of the past. So the undermining does not depend on what higher-order beliefs the agent happens to have about her own beliefs and desires. Correlations to the past are undermined, even when agents lack full self-knowledge. A similar response

[123] But for responses, see Eells (1982; 1984; 1985) and Jackson and Pargetter (1983).

applies to the case of the observing agent. If Gina is properly deliberating, her decisions for the sake of past states aren't evidence of past states. This undermining doesn't depend on Gina or any observer, having any particular beliefs about Gina's beliefs or desires.

Finally, it might seem that undermining will not occur if the mediating states aren't present at the start of deliberation, but occur at later stages. If Gina's initial beliefs and desires aren't evidence of the past and it is only the beliefs that she comes through deliberating that provide good evidence, it might seem evidential decision theory will provide the wrong advice. In response to this and other more complex cases, Eels (1984) develops a *dynamic tickle defence*—for discussion, see Horgan ([1981]), Price (1986; 1991; 1993) and Horwich (1987, ch. 11). Having made an initial judgment of the expected utility of an act, the agent doesn't immediately decide, but returns to deliberate with this judgment as an additional input. This process then iterates. At any stage in the process the *judgment* may screen off the relevance of the decision or act for the past state. Evidential decision theory only issues a recommendation when the iteration produces no change in the expected utility of the acts. This will be after the screening off has occurred, provided the screening off occurs, and a recommendation is issued at all.

But such a response is not needed to defend the deliberative account of causation. Unlike evidential decision theory, the deliberative account does not require there to be a particular point in deliberation by which screening off occurs and a decision is recommended. What is required is that when the agent *decides*, the deliberation is proper and the decision has the right evidential relations. This *deciding* necessarily occurs at the conclusion of deliberation. So any complex adjustments to the agent's beliefs and desires are necessarily taken into account.[124]

For related reasons, the deliberative account also avoids problems raised by Ahmed (2005) against Price's version of the tickle defence (1986, 1991, 1993). These problems arise when undermining or screening off is appealed to in defence of evidential decision theory. Price argues that if Gina decides not to smoke *because* she thinks not smoking is evidence of not having the gene, this evidential reasoning provides a competing causal explanation for why she decides not to smoke. So her decision will not in fact be good evidence for not having the gene, and Gina loses her evidential grounds for making the decision. A decision not to smoke on this basis is 'unstable'. Price argues that evidential decision theory should require us to reach equilibrium judgements, where the evidential relevance of our decision is not undermined by a judgment that we will decide on its basis. The only stable positions are where the agent either takes her decision to be probabilistically irrelevant to past states, or suspends judgment on the matter (1986, p. 201). So agents won't be led to decide for the sake of the past. However, as Ahmed (2005) argues, there are cases where evidential decision theory, if required to reach an equilibrium judgement, will reach no recommendation, or will give the wrong advice, even by evidential lights. These are problems for defending evidential decision theory. But not for defending a deliberative account of causation. A deliberative account does not have to issue recommendations. To explain causal asymmetry, it is enough that, were an agent to decide for the sake of the past, the correlation would be undermined.

[124] Jeffrey (1981; 1983) uses a similar response to defend evidential decision theory, by considering an agent's attitude towards his decision when the decision is made but not yet enacted. But this 'ratificationist' solution faces problems of its own as a defence of evidential decision theory—see Horwich (1987, pp. 188–9).

There are further concerns to be faced in defending a deliberative account of causation. If agents are never able to decide based on beliefs that are meant to track evidential relations, then a causal relation that corresponds to evidential relations will not be so useful. But we should keep in mind that evidential decision theory does remain a strong candidate for many ordinary decision situations. A more serious concern is that the evidential standards the deliberative account appeals to are much more stringent than what agents actually use. Perhaps merely having a desire does not provide Gina with evidence of the gene, under an ordinary notion of evidence. Appealing to overly high standards would place unreasonable epistemic demands on agents. And it would also risk making all deliberation improper. If merely having beliefs and desires evidentially settles what Gina decides, it seems Gina will never be in a position to properly deliberate, even on the *future*. A similar concern is raised against the screening off defence of evidential decision theory. Skyrms (1980, p. 131), Lewis (1981, p. 10 n.8) and Price (1986, pp. 204–6) all argue that if an agent has evidence of her beliefs and desires, and believes she will choose in accordance with them, as the tickle defence standardly requires, deliberation is closed off, even in ordinary future-directed cases. The (subjective) evidence always already settles what decision the agent will make, and she cannot reasonably deliberate.

To fully respond to such concerns, a defence of the deliberative account would need to show that the standards appealed are sufficiently weak, and do not take agents far beyond what they can reasonable access and use. But while I won't attempt to answer these concerns in full here, let me point towards two considerations that suggest a sufficiently weak notion can be found. Firstly, even if the careful accounting of an agent's complete set of mental states and the relevant parts of the external environment could lead one to predict her decision, it is often unreasonable to suppose that any agent could make these calculations. If so, many actual cases of deliberation will remain proper. Unless we think that agents are systematically insensitive to their evidence, the epistemic model implies that in many ordinary cases, an agent's evidence does not settle what she will decide and do.

Secondly, deliberation is an iterative process—an agent's beliefs and desires while deliberating can be further inputs to deliberation. This is something that the dynamic tickle defence rightly drew our attention to. Because deliberators are complex and self-reflective, a deliberative state, even one that provides evidence of the past, is often not enough to settle what an agent decides. So deliberation *can* on the future can be proper. For example, Gina may start out her deliberation with a desire to smoke, a desire that strengthens as she considers the charms of smoking, until her having the gene is evidentially settled. But her deliberation does not stop there. Her beliefs and desires can change from this point, in ways not predictable from her earlier states, for a reasonable standard of evidence. She might bitterly recant her desire to smoke and decide not to smoke. What she finally decides is not settled by her earlier desire (see figure 14). So she can properly deliberate in many ordinary future-directed cases, even if she still can't properly deliberate for the sake of the past.

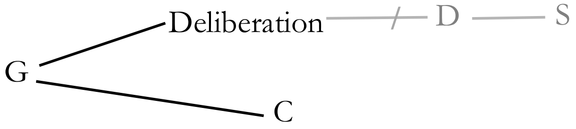


Figure 14: The evidence Gina has while deliberating on smoking (S), given the causal structure in figure 12. Even if Gina's evidence in deliberation settles her having the gene (G) and getting cancer (C), it doesn't settle what decision she'll make (D). So she can properly deliberate on the future.

Note that this defence of the deliberative account does require a careful balance—allowing that there are ordinary cases of proper deliberation, but none that imply backwards causation.

The deliberative account does not imply backwards causation. If decisions are reliably correlated with past states of affairs, this goes via agents having evidence of these past states, due to observations, other external evidence, or deliberative states like beliefs and desires. This evidence makes deliberation on the past improper. It is only when these mediating states are absent that deliberation is proper. But in this case we won't expect decisions to be evidence of the past states. So in neither case is backwards causation implied. The situation towards the future is very different. Correlations between decisions and states further in the future are not mediated by deliberative states. Deliberative states do not provide evidence of future states that is independent of what decision an agent makes (except in the case of spurious correlations). This means that when deliberation on the future is proper, there can still be evidential correlations between decisions and future states. So there can be forwards causation. Putting these two pieces together, causal asymmetry is due to the fact that correlations between decisions and the past are mediated by deliberative states, while correlations towards the future are not.

## 5.4 The Temporal Asymmetry of Deliberation

To explain causal asymmetry, I appealed to the fact that deliberation undermines correlations to the past, but not (directly) to the future. Deliberation only undermines correlations to the future by way of the past (in cases of spurious correlations). I explained this undermining by appealing to the fact that only evidential relations between decisions and past states are *mediated* by deliberative states. Correlations to the future are not mediated by states that are part of deliberation. In this section I explain this why this mediation is asymmetric by appealing to the fact that deliberation comes *before* deciding. To explain this asymmetry in turn, I'll appeal to an asymmetry in our epistemic access to the world. It is because we have records of the past but not the future that we deliberate before deciding. Statistical-mechanical accounts also appeal to this epistemic asymmetry to explain causal asymmetry. But the epistemic asymmetry will play an importantly different role in each explanation.

### 5.4a Explaining the Asymmetry of Undermining

Before I begin, let me consider a competing explanation for the asymmetry of undermining. Price uses this explanation to defend his preferred version of evidential decision theory. Recall, Price argues that Gina should decide to smoke, because if she decides *not* to smoke in order to avoid having the gene, the correlation between smoking and the gene is undermined. Price claims this undermining occurs because Gina's deliberation provides an 'alternative *causal* explanation' for why Gina smokes, one that interrupts the evidential relation that would otherwise obtain between the gene and her smoking (see figure 15) (1991, p. 166, my emphasis; see also 1986, p. 201; 1993, p. 261; 2007, pp. 281–2).

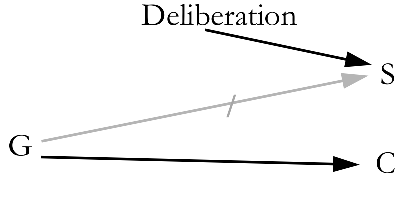


Figure 15: Price's explanation of the asymmetry of undermining. Gina's deliberation provides a competing causal explanation for her smoking (S), which undermines the evidential relation that would otherwise obtain between having the gene (G), or getting cancer (C), and smoking.

The reason only correlations to the past are directly undermined in this way is because deliberation comes temporally prior to acting. Deliberation doesn't undermine causal relations to the future (except by way of the past). Ultimately, causal asymmetry is then due to deliberation coming before action.

But there are two problems with Price's explanation. Firstly, he relies on an unexplained asymmetry—that deliberation precedes action. Price does not attempt to explain this asymmetry, and doesn't take offering such an explanation to be part of the philosophical project of explaining of causal asymmetry. His explanation only begins from our temporal orientation as agents (1993, pp. 261−2; Price and Weslake 2009, p. 434). While he accepts that science *may* explain our deliberative orientation, (ibid., p. 434 n. 23), he doesn't suggest how this might be done without presupposing causal asymmetry—as much of higher-level science does. More often, Price suggests treating deliberative asymmetry as an 'ultimate contingency' (1993, p. 260). Price's stance on this issue makes sense, given his 'perspectivalism'. Price's project is to explain away apparently objective features of the world as arising merely from features of our perspective. For this purpose, we don't need to know how the features of our perspective arise. My project is different. I want to account for causation in a way that makes sense of its relevance for agents, but still relates causation to objective features of the world.

A second more serious concern with Price's explanation is that it is explicitly causal. Undermining occurs when an agent decides *on the basis* of a belief in a correlation, not merely when she believes the correlation holds. It is 'in virtue of [its] causal role that the judgement is self-defeating in the way described' (1986, p. 201, see also 2007, p. 282). Price's appeal to causation is not a problem for defending evidential decision theory—evidence may be based on causal assumptions. But this appeal to causation is a problem for explaining causal asymmetry. If the fact that causes come before their effects is needed to explain why causes comes before their effects, we have a viciously circular explanation. This is not how Price sees the situation. He claims that his account leaves 'no asymmetric modal notion left unaccounted for' (1993, pp. 261−2), and that the probabilities appealed to are 'not dependent on a prior modal notion' (2007, p. 281). But in appealing to causal features of deliberation, we do rely on prior modal notions.

To explain undermining, while avoiding Price's appeal to causation, I suggested the following. In cases where an agent's decision are evidence for external states of affairs, this goes *via* the agent being in a particular state at a time in between. It is only when this mediating state is present that we expect the evidential relation to hold. I argued that relations to the past are mediated by deliberative states that form part of deliberation, such as beliefs and desires. Relations to the future are mediated instead by decisions, intentions, or intentional acts. Both these types of mediations between an agent's decisions and her environment are instances of a more general phenomenon: that evidential relations between states of affairs at two different times are mediated by states at times in between. While we might not know these mediating states, and may reason directly from states at different times, if we look closely at the features of the situation, we expect to find a continuous series of evidential relations involved.[125] In the common causes cases above (figure 12, p. 224 and

[125] This evidential mediation does relate to a causal phenomenon: that causal relations, at least at our world, are continuous through time. If two states of affairs at different times are causally related, we expect to find a series of states that occur at times in between that are also causally related. Causal continuity could even explain evidential mediation. But as with evidential relations more generally, the fact that they can be explained in causal terms does not mean they can't be appealed to in evidential terms.

figure 13, p. 226), for example, we expect there to be mediating states at times between Gina having the gene and getting cancer, that are evidentially tied to both these states. This is the evidential mediation that we need to explain undermining.

Using evidential mediation, and the fact that deliberation comes before decision, we can explain why undermining is asymmetric. If deliberation comes before decision, and not after, only correlations to states of affairs further in the past are mediated by deliberative states. So only correlations to the past are undermined. If Gina's deliberation for the sake of the past is proper, her decision now won't be evidence of past states. But correlations between decisions and future states are not mediated by deliberative states. So Gina's deliberation may remain proper, even while her decisions are evidence of future states. So evidential relations to the future are not undermined by deliberation. If we can explain why deliberation comes prior to decision, we will have explained asymmetric undermining without relying on causal asymmetry.

Furthermore, even if we explain undermining in a different way, we should still expect to appeal to the fact that deliberation comes before deciding. This is certainly the case with existing explanations, including those by Price (1993; 2007), Woodward (2003) and Pearl (2000). Woodward and Pearl effectively treat actions and decisions as interventions, where the deliberation breaks causal relation further upstream. Insofar as these accounts explain asymmetric undermining, they rely on deliberation coming not only causally prior, but also temporally prior to decision. The situation is similar if one appeals to Reichenbach's common cause principle (1956)—roughly the fact that correlated events must have a common cause but not necessarily a joint effect. Unless we appeal to temporal assumptions about deliberation, we will not be able to explain why only correlations to the past are undermined by deliberation. Whether one uses an evidential or causal route to explain asymmetric undermining, it is crucial that deliberation comes before decision.

### 5.4b Why Deliberation Precedes Decision

To explain asymmetric undermining, I appealed to the fact that deliberation precedes decision. But why does this asymmetry hold? In this section, I explain this deliberative asymmetry using the epistemic model of deliberation from chapter 2 and an epistemic asymmetry—that we have memories of the past, but nothing like memories of the future.

Begin with the fact that we have memories of the past but nothing like memories of the future. This means that in ordinary cases of deliberating, we are certain about the immediate past in our vicinity, independently of what decision we make. As Tamsin deliberates on whether to take her umbrella, for example, she is certain she was walking through the hall a minute ago. And she is certain of this, independently of what decision she makes now. We also have knowledge of our immediate futures. Tamsin may be certain she'll leave the apartment in the next two minutes and that the hall will be in its usual place. But we typically have knowledge of the future either by having general beliefs about the present (or past) and its reliable behaviour, or by having knowledge of our current decisions, intentions and habits and their reliability. Tamsin's certainty of her own immediate past is not dependent in this way—independently of her decisions, intentions, habits or beliefs about the present, she has *memories* of her past. She does not have anything like memories of her future.

According to the epistemic model of deliberation, an agent can't deliberate if she is certain of what decision she'll make. Doing so prevents deliberation serving its usual epistemic function. So if Tamsin has memories of her past decisions, she can't deliberate after deciding. But because Tamsin does not have anything like memories of her future, she can deliberate *before* deciding. As she deliberates, her future decision remains suitably uncertain. As far as an agent's deliberative structure is generic, and tracks general epistemic features, the epistemic asymmetry leads to a temporal asymmetry of deliberation.[126]

Note that this explanation of a deliberative asymmetry is not available on the competing accounts of deliberation considered in chapter 3, including those by Price (2012) and Ismael (2007). These accounts deny that there are ignorance conditions on deliberation. So the fact that an agent has memories of her decisions does not prevent her deliberating *after* deciding. This gives us an additional reason to prefer an epistemic account of deliberation in which an agent must be ignorant of her decision as she deliberates.

One might be concerned that memory is insufficient to rule out deliberation on the past. Memory functions reliably only over a limited range. Some of the decisions you made yesterday, for example, you have no doubt already forgotten. Can you then deliberate on them? The problem is your 'deliberation' today would not have an evidential bearing on your original decision. The evidential relation between your deliberation and the decision is broken by your not having evidence of your decision at times in between, as is required by the setup. While you might 'deliberate' in some attenuated sense, your deliberation would

[126] Memories also allow agents to pick up on correlations between decisions and actions, satisfying another of the epistemic requirements argued for in chapter 2—that agents take their decisions to be good evidence of their results.

not serve the epistemic function argued for in chapter 2—it would not be a way of settling what you decide and do.

There are also general temporally *symmetric* features of deliberation that contribute to us remaining uncertain of our future decisions. Firstly, we can't observe the actual physical mechanisms that determine how we decide. Our neural wiring is small, complicated and hidden from view. Secondly, the correlations between states of the world we respond to and our decisions are often complex and difficult to keep track of. While there is some reliability in how we decide, humans aren't simple creatures, and respond to a range of factors in a variety of different ways. It is not easy to model how we will decide on any one occasion. For example, consider whether a fireman can take the heat of the floor as evidence he will decide to leave a burning building—an example from Holton (2006). In order to make an accurate inference, he must take into account the nature of his values and desires, the way he balances the risks involved against the possibility of saving the building, and the significance of all this evidence. His decision will not be amenable to simple modelling. Thirdly, features of past deliberations may be relevant to how we later decide. We are self-reflective creatures, able to use our previous deliberations as further inputs to our current deliberation. The fireman, for example, may be more reluctant to risk his team, if he did so previously. Because of these three factors, our decisions are not easy to predict. And so an agent can be suitably ignorant of the future so as to deliberate. But memory prevents us being suitably ignorant of the past.

We can imagine more complex creatures, where memory or other recording faculties are incomplete in certain systematic ways, or sometimes reverse direction. According to the epistemic model, such creatures could sometimes properly deliberate on the past, and sometimes properly deliberate on the future. But our cognitive architecture is not like this. And so our deliberation has only one temporal orientation. Nor is this a surprising feature. As I'll argue shortly, memory is an example of a more pervasive epistemic asymmetry that statistical-mechanical accounts rightly draw our attention to.

### 5.4c Remembering the Past

So far my explanation of causal asymmetry has relied on an epistemic asymmetry—that we have memories of the past, but nothing like memories of the future. To complete the explanation of causal asymmetry, I now consider why this epistemic asymmetry holds. I do so by appealing to an initial low-entropy state of the universe. Albert (2000; 2015) and Loewer (2007) also use a particular low-entropy state, the Past Hypothesis, to explain an asymmetry of records, on the way to giving statistical-mechanical accounts of causal asymmetry. But these statistical-mechanical accounts appeal to entropy in a different way from the deliberative account—one that makes these accounts harder to defend.

In section 5.2c, I argued that Albert hadn't shown that the Past Hypothesis is sufficient to explain our reasoning to the past using records. While we might need an initial low entropy state to rule out conspiratorial worlds, this doesn't show that a Past Hypothesis (plus the dynamical laws and statistical postulate) is sufficient for records of the past. And it doesn't show that the Past Hypothesis is necessary either. All that's required to rule out conspiratorial worlds is a low entropy initial state (a 'past state'). To defend a statistical-mechanical account of causation, Albert and Loewer need to identify necessary and sufficient conditions on reliable records of the past—they need an algorithm for evaluating counterfactuals that picks up on how we infer to the past and future. But if what we want to do is *explain* why we have memories of the past and not the future, all we need is to identify the relevant *difference maker*—a feature that allows for records of the past, but not records of the future.[127] And we have that, in there being a low entropy constraint on the past, but no similar constraint on the future. The arguments I considered above all concerned whether Albert had identified necessary and sufficient conditions on our normal procedures for inferring to the past. None of these arguments attacked the claim that a low-entropy boundary condition is a necessary condition on records.

How do records work? Recall, records are local states of affairs in the present that allow us to reliably infer to states of affairs at other times, in ways that go beyond what we can infer to using the dynamical laws and statistical postulate. Records take us beyond what we can infer to by assuming generic behaviour towards the past and future, or tracing what happens to the whole system over a period of time, using dynamical laws. For example, say I have a memory of being given a toy dog, 'Penny', on my 5th birthday.[128] Given I have this memory now, I can infer that I was indeed given Penny all those years ago. I can infer this, without knowing that children are usually given toy dogs on their 5th birthdays, or without knowing what has happened to Penny between then and now. I may also infer that my nephew will receive a toy turtle 'Bruce' on his 5th birthday. But I infer this, either because he likes turtles, and children tend to get what they like for birthdays, or because I believe that there's a turtle in the store, that I will buy it, keep it and give it to him—I know what will happen to the turtle between now and then. We use records to reason to the past, not the future.

[127] The idea of difference-making underlies prominent accounts of causal explanation (Woodward 2003; Strevens 2008), as well as counterfactual accounts of causation more generally (Lewis 1973a). Strevens, in addition, argues that difference-making extends naturally to domains outside the causal, including mathematics. Here I use difference-making in giving a non-causal scientific explanation.

[128] There is also a question of how memories come to have the content they do—why my memory is a memory 'of' this event. This may involve causal factors. However that story proceeds, we nevertheless have good reason to believe that memories will be reliable indicators of past events.

A low-entropy boundary condition is a necessary condition on records. Only with a low-entropy boundary condition can there be reliable forms of inference by physical means that takes us beyond what we could infer to using the dynamical laws and the statistical postulate.[129] Recall, a low-entropy boundary condition is simply a restriction on the phase space that is occupied by the universe at one temporal end. If the restriction goes beyond what is implied by the universe's present macrostate, the dynamical laws and statistical postulate, it allows for a form of inferring from the present that goes beyond these as well. And it is only with such a restriction that we can infer in ways that goes beyond these.

Given my aim in this chapter is to explain causal asymmetry, I won't argue for a particular account of how we infer to the past using records like memories.[130] It is sufficient for my purposes that reliable asymmetric forms of reasoning like memory are an instance of the more general phenomena of records, and that a necessary condition on records is a low-entropy boundary state. Whether Albert's explanation or some other accounts for how we reason using records, a crucial feature of these explanations is their appeal to a low-entropy condition. The fact that there is a low-entropy condition on the past, but not the future, is what explains why there can be records of the past and not the future. And it is records, like memories, that give us reliable knowledge of past states of systems beyond generic knowledge and independently of tracing out the state of the system between now and then. It is this kind of knowledge of our actions, decisions and further states that prevents deliberation. And so it is the general fact that we have records of the past, but not the future, that explains why we deliberate towards the future, and not the past. Memory remains a

[129] Noting, as earlier, that the low-entropy boundary condition need not be at the start of the universe.

[130] For possibilities, see Albert (2000, ch. 6; 2015, ch. 2), Reichenbach ([1956], chs. 3–4), Grünbaum (1963) and Smart (1967).

particularly important form of reasoning using records because it is both common and reasonably reliable concerning the recent past. These features are what make it particularly well-placed to prevent deliberation after deciding. But what more generally explains why we deliberate before deciding is the fact that we have records of the past, but not the future. And this is due to a low-entropy past state.

## Conclusion

Causes come prior in time to their effects because of features of deliberation. It is because we deliberate before deciding that evidential correlations towards the past are undermined by deliberation. Because causation corresponds with the evidential relations we use in deliberation, causation is not directed towards the past. Why are correlations towards the past undermined? Because we deliberate *before* we decide. Why do we deliberate before we decide? Because we're suitably ignorant of the future, but not the past. We have records of the past, evidence of the past that is independent of what we decide now. Why do we have records of the past and not the future? Because there is a low-entropy boundary condition on the past and not the future. Figure 16 summarises this explanatory structure.

The deliberative account traces causal asymmetry back to a low-entropy initial condition—the same type of feature appealed to in statistical-accounts by Loewer and Albert. But while statistical-mechanical accounts aim to reduce causation to other regularities, and so must identify necessary and sufficient on reasoning to the past and future, the deliberative account does not. Instead, the deliberative account needs to explain why evidential correlations to the past are undermined by deliberation, but correlations towards the future are not. And it can do this without identifying necessary and sufficient conditions on inferring.

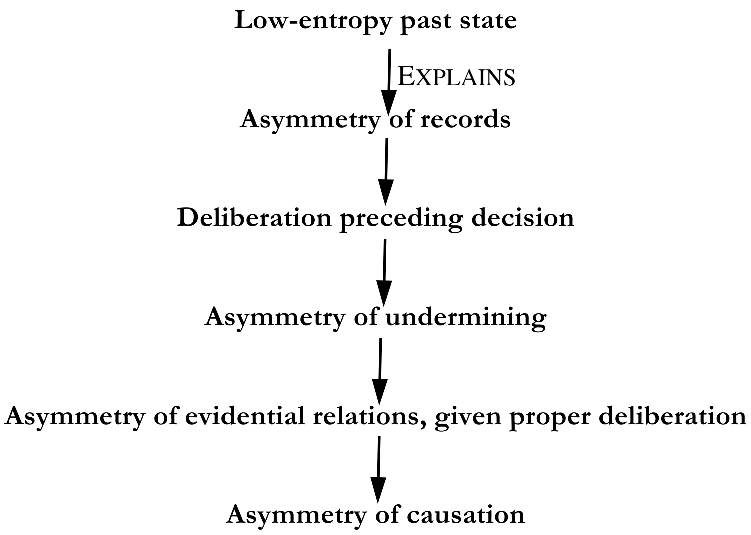


Figure 16: How the deliberative account explains causal asymmetry.

The deliberative account does not aim to reduce causal relations. Instead, it aims to illuminate the general explanatory and justificatory structure that underlies any successful account of causation. And it identifies, at the most general level, how the evidential structure of the world relates to its causal structure.

# Chapter 6

# Conclusion: Our Place in Science

## 6.1 Making Sense of Causation

We should understand how causation fits into the world by thinking about its relevance for deliberation. It is because we're agents, reasonably able to deliberate and decide on one thing in order to achieve another, that we pick out certain relations as causal. Causal relations in fact correspond to the evidential relations that hold between our decisions and outcomes we seek. For example, Tamsin's taking her umbrella counts as a cause of her staying dry just in case her deciding to take her umbrella is evidence of her staying dry. Agents like Tamsin can therefore use their knowledge of causal structure to make decisions that are evidence of the outcome they seek. And this is why causal relations should matter to us.

What does this deliberative account imply about causation? Firstly, the account supports non-fundamentalism about causation, where causation is not part of the fundamental metaphysical structure of the world. Non-fundamentalism might be true either because causation reduces to more fundamental relations or because the world has no metaphysically fundamental level. One reason for this support for non-fundamentalism can be found in the foundations of the project. In chapter 1, I motivated there being a problem with causation by considering Russell's challenge: fundamental physical theories don't mention causes. Insofar as we want our fundamental metaphysics to be guided by fundamental physics, this is a *prima facie* reason not to take causation to be fundamental. The deliberative account further supports non-fundamentalism by showing what role causal relations play, and why we should pick out certain relations as causal, even if causes are not needed to push and pull things around at the fundamental metaphysical level. I argued that we need causal relations for deliberation. The deliberative account explains why causal relations matter, even if they're not fundamental. While the account is compatible with causation being fundamental, it provides more support for non-fundamentalism.

There is another sense in which the deliberative account supports non-fundamentalism about causation. The deliberative account cannot be applied to system-wide causation. This is a feature it shares in common with interventionist accounts. The deliberative account relies on a separation between the deliberating agent (or intervener) and the rest of the system, and only directly derives causal relations between states that don't include the deliberating agent (or intervener). If causation were fundamental, there would be no need for such a restriction. Fundamentalist views, moreover, have tended to take the whole state of a system at one time to cause its state at the next time. Excluding system-wide causation is in keeping with a broadly Russellian program. Russell's challenge shows that system-wide causation is not needed for the state of affairs of the system at one time to determine its state at other times—non-causal fundamental physical laws can do that work.

The deliberative account also supports non-fundamentalism about causation because it takes type-level causation to be more primary than token-level (or 'actual') causation. If causal relations were part of the fundamental structure of the world, needed to push and pull things around, this would suggest that token-level causation would be the more primary phenomenon—causal relations would capture the 'force-like' relations that hold between actual events. If causal relations correspond to the evidential relations of use to deliberating agents, this suggests type-level causation is the more primary phenomenon. Evidential reasoning is reasoning we often engage in hypothetically—we think about what would be evidence for what, even if neither the evidence nor the state obtains. Deliberation is also a hypothetical exercise—we are interested it what would follow from our decisions, knowing we will only decide in one way. Both kinds of reasoning require us to identify similar cases or types, suggesting causation is also primarily type-based. The deliberative account, furthermore, includes an additional counterfactual layer—it appeals what evidential relations would hold, were an agent to properly deliberate. I suggested that these counterfactuals should also be evaluated in a type-based fashion, by appealing to non-causal similarities between systems and laws. For these reasons, while the deliberative account is strictly compatible with taking causation to be token-level (insofar as evidential relations can also be token-level) it supports a type-based approach—in keeping with non-fundamentalism.

Finally, the deliberative account also supports non-fundamentalism about causation by providing important resources for reductive accounts. Firstly, the deliberative account can justify why the relations picked out by a reductive account deserve to be called causal. If a reductive account reasonably satisfies both sides of the evidential biconditional, relating evidence to causation, a defender can appeal to the relevance of causation for good decision-making to justify their account. Secondly, as I argued in chapter 4, the deliberative account can be used as a half-way step to ultimately reduce causation to fundamental physical laws and contingent features. If evidential relations are objective probabilities, and objective probabilities reduce to laws and contingent features, as Lewis, Loewer, Albert and others intend, then the deliberative account supports a reduction of causation to fundamental laws and contingent features. If this reduction works out, we have a particular strong answer to Russell's challenge—we have a direct relation between causation and fundamental physical laws.

However, the deliberative account doesn't imply that causation is reducible. While I have emphasised how the deliberative account is relevant for reductive accounts, it is also relevant for non-reductive interventionist accounts of causation. Woodward (2003) and Pearl (2000), for example, don't explain why causation is useful or why causes come before their effects. But they do derive evidential relations from causal relations. This means they can appeal to the deliberative account to explain why interventionist counterfactuals are useful—they are the evidential relations of use to deliberating agents. Interventionist accounts can also appeal to deliberation to explain why we select exogenous and endogenous variables as we do, why agents approximate interveners and why intervention breaks arrows temporally upstream—it is because deliberation effectively severs evidential correlations towards the past that deliberators approximate interveners, and so why causes comes prior to their effects, at least in deliberative contexts. While interventionist accounts and the deliberative account will likely disagree over particular cases, they remain broadly compatible, and interventionist can still learn important lessons from the deliberative accounts.

However these further debates about reduction work out, the deliberative account remains informative and explanatory. It remains so even if, quite generally, metaphysical notions of reduction turn out to be ill-formed. The deliberative account relates causation to evidential relations and deliberation. We have a sufficiently independent grasp of each of these for the deliberative account to remain interesting, even if causation is not reducible to evidential relations. In chapter 2, I presented an epistemic model of deliberation without appealing to causal notions. In chapter 4, I argued that evidence could be characterised by its normative role in inferential reasoning. Because we have sufficient grasp of deliberation and evidence independently of causation, the evidential biconditional remains substantive and contentful—even while it does not settle whether causation is reducible. The deliberative account in fact exemplifies a general metaphysical project of connecting various relations to one another in interesting and explanatory ways, without taking some relations to be more metaphysically fundamental than others.

If causation is not ultimately reducible, does that mean the deliberative account no longer relates causation to fundamental laws? Can we no longer answer Russell's challenge? Not necessarily. One can relate evidential relations and objective probabilities (and thereby causation) to fundamental laws, without taking these relations to be reductions—without taking fundamental physical laws to be *metaphysically* more fundamental than evidential relations. One way to do this is to make use of the explanatory resources within science. Fundamental physics aims to *explain* the success of higher-level sciences, including non-fundamental physics. If evidential relations and objective probabilities are understood by reference to the role they play in higher-level science, scientific explanations can relate fundamental laws to evidential relations and objective probabilities. The deliberative account would then relate fundamental laws and causation—without implying metaphysical reduction. I leave exploring the relation between evidential relations and fundamental laws for future work.

I've argued that the deliberative account supports non-fundamentalism about causation. Does this mean that the deliberative account is thereby anti-realist about causation? No. A relation can be real without being fundamental.[131] The deliberative account takes causation to

[131] One could of course *define* realism as implying fundamentalism. Then the deliberative account would support anti-realism, while being compatible with realism. But such definitions tend to do a poor job of capturing views like my own.

be just as real as other important relations like laws and objective probabilities. By appealing to how these relate to one another, the deliberative account aims to give us a better understanding of what causation actually is. Nor does the account imply eliminitivism about causation: that causation is scientifically or metaphysically suspect, or that we should do away with talk of it. While Russell argued for a wholesale rejection of causal notions, that has not been my project. The account also does not imply that causal talk is non-representational, or non-truth-apt, or otherwise contrasts with ordinary descriptive discourse.

There's another way in which the deliberative account might seem to imply anti-realism about causation. Because the account appeals to agency, it might seem to imply that causation is mind-dependent, or otherwise depends on features of agency—such as our agential perspectives. But the deliberative account does not imply that causation is mind-dependent, or perspectival. As I argued in chapter 4, causal relations do not turn out to depend on our beliefs, desires, practical abilities or perspectives. Instead, the evidential biconditional ensures that causal relations are at least as objective as the evidential relations we use when we reason. If we take evidential relations to be objective probabilities, or relate them to laws, the objectivity of causation is further secured.

The deliberative account instead supports the following picture. Understanding the relevance of causation requires appealing to agency and deliberation. Even though causation might have seemed like a deep fundamental feature of the world, something whose importance we could make sense of independently of features of ourselves, we should make sense of it by reference to us. Causation matters to us because we are agents who can reasonably decide on one thing in order to achieve another. While causal relations appear from outside the perspective of the deliberating agent, they are peculiarly relevant for deliberators—and this is why conceive of the world in causal terms. There's a surprising way in which an objective and scientific relation (causation) turns out to relate closely to our interests as agents.

This is to take broadly pragmatist or idealist approach to causation. The account does not rule out metaphysical accounts of causation. We can take causation to be fundamental, otherwise not reducible, or reduce to other regularities. But the account itself makes none of these commitments. The explanatory power of the deliberative account derives from the fact that it identifies how broad-scale evidential and causal features of our world relate—and making clear how deliberative agency is essential in understanding causation.

Similar lessons hold regarding the asymmetry of causation—why causes come before their effects. In chapter 5, I argued that causal asymmetry derives from a temporal asymmetry of deliberation—the fact that we deliberate before we decide. This means correlations towards the past, and not the future, are undermined by deliberation. The direction of causation is therefore contingent—causes could have come *after* their effects (rigidifying on the actual direction of time). The temporal asymmetry of causation is not an unexplainable primitive. But the asymmetry of causation is not thereby mind-dependent or perspectival. Nor is it the case that causation could have 'easily' gone the other way. Even though agency is essential to understanding causal asymmetry, causal asymmetry still depends on other physical asymmetries as well. This is in part because the physical structure of the world *determines* our orientation as agents, including the fact that we deliberate before we decide. Agents are physical beings, whose features we can expect to explain scientifically. In chapter 5, I traced the asymmetry of deliberation back to an epistemic asymmetry—the fact that we have records of the past, but not the future. This epistemic asymmetry is not up to us or mind-dependent in any substantive sense. Furthermore, it is an asymmetry we can expect to make sense of in terms of other physical asymmetries, particularly asymmetries relating to entropy (Albert 2000). Even though agency is essential to understanding the temporal asymmetry of causation, causal asymmetry is ultimately due to objective asymmetries in the physical world.

There are limits to the deliberative account of causation. It only applies in cases where we can make sense of what an agent's evidence would be, were she to be properly deliberating. We may have difficulty working out what evidential relations would hold for a deliberating agent, particularly in conditions that don't allow for agents—such as at the early universe. Reductive accounts may play a useful role in such cases, and help extend causal concepts from cases we do ordinarily encounter to those we do not. We'll also have trouble deriving causal asymmetry if we consider very simple universes, with no entropic structure. If an agent were in such a universe, would they deliberate before deciding, or the other way around? My project has focussed instead on the relevance of causation for decision-making in the ordinary cases we actually encounter. While I've suggested how the account can be extended to cover causal relations where agents aren't able to properly deliberate (chapter 4), recovering our intuitive causal judgments in such cases is not my main concern. I'm more interested in why causation matters to us, and so how causation operates in cases we do actually encounter. First-order metaphysical accounts of causation may do better at recovering our causal intuitions. But the deliberative account does better at uncovering the broad-scale explanatory and justificatory structure that underlies these accounts.

### 6.2 Making Sense of Agency

I began this project with a second motivation—to understand how free agency fits into the world, given the picture presented to us by science. More particularly, I aimed to explain our apparent ability to deliberate and decide on different options, given that fundamental laws and previous states (effectively) determine what we do. The epistemic model of deliberation I put forward in chapter 2 answers this aim. The model characterises deliberation by placing conditions on the beliefs we can and can't reasonably have while deliberating. I argued, for example, that we must be uncertain of our decisions, and yet certain our decisions settle a state of the world. This model explains our apparent freedom in deliberation, by appealing to a particular combination of certainty and uncertainty.

In chapter 3, I defended the model against competing epistemic accounts—particularly those that explain our apparent freedom in deliberation by appealing to our 'license' or justification to believe in different ways about what we'll decide or do. I argued that, despite appearances, these accounts don't actually explain our apparent freedom in epistemic terms. Instead, they rely crucially on desire-responsiveness or primitive notions of freedom. I also argued against perspectival accounts that separate practical and theoretical reason into two different stances. Such approaches can't explain causation and causal asymmetry without introducing further unexplained asymmetries. While my main aims in these chapters was to present and defend the epistemic model, I also addressed more general questions about what options are available to an agent and how beliefs evolve during deliberation—something of importance to the foundations of decision theory and philosophy of action.

The epistemic model of deliberation I defended is deflationary in a number of ways. Firstly, it privileges the epistemic function of deliberation over its casual and normative functions. It takes deliberation and deciding to be ways of settling what *will* be the case, not primarily ways of settling what *ought* to be the case, or ways of *bringing about* states of affairs. Secondly, the epistemic model doesn't imply that agents have primitive metaphysical powers, such as the ability to decide or determine states in ways not determined by previous states, causes and laws. Thirdly, the account, as I've employed it, explains why we *appear* free in deliberation—not why we *are* free.

These deflationary features might suggest that that the epistemic model is anti-realist, and that agency and freedom are merely epiphenomenal, and not real features of us or the world. Again, this is not what the model implies. While I've focussed on explaining why we *appear* free, this doesn't, by itself, give us a reason to think we're *not* free—not without presupposing something further about what such freedom must amount to. The model identifies structural conditions that are required for deliberation of any reasonable type. These epistemic conditions on deliberation are often satisfied. We are often ignorant of what we will do, yet certain that our decision settles what we will do. In this sense, deliberation is perfectly reasonable. Given existing limits to our beliefs about the world, there is a place for deliberation.

The epistemic model does, however, undercut arguments from our apparent freedom in deliberation to more robust conceptions of freedom. If we can make sense of apparent freedom without appealing to robust normative facts, for example, then our apparent freedom in deliberation gives us no reason to commit to such facts. But, again, this result only supports anti-realism if we were already to committed to deliberation involving robust normative facts. Without introducing further standards by which deliberation, epistemically conceived, counts as less than real, the epistemic model does not support anti-realism.

The epistemic model can even be used to give a compatibilist account of free agency. As noted in chapter 2, epistemic ignorance-based models have explicitly defended as compatibilist accounts of freedom (Kapitan 1989). The conditions of the epistemic model can be satisfied, even though the world operates according to laws and causes, making the epistemic model broadly compatibilist. Because the model is compatibilist, it also provides a strong answer to the second motivation of the project: to understand how agency fits into the world as presented by science. The epistemic structural features required for deliberation can be satisfied, given that we live in a world that operates according to laws and causes, and given our epistemic limitations as agents.

There is also a sense in which my accounts of deliberation and causation imply we should take agency *more* seriously as part of our picture of the world. I've argued that causation can't be understood independently of deliberation. Causal relations correspond to the evidential relations of use to deliberating agents. An epistemic account of deliberation is therefore not only compatible with a world that operates according to laws and causes—it's required to make sense of that world. Understanding agency turns out to be bound up with understanding something of deep importance to our scientific picture.

If agency is required to make sense of causation, then this project provides a further kind of reconciliation between science and agency. It may have seemed that the causal structure of the world threatened to leave agents out of the picture. I've argued that we can only make sense of causation by thinking about agency. Seeing the relevance of agency for causation may help reconcile us to viewing the world causally.

### 6.3 Expanding the View

I've argued that we should make sense of causation by way of deliberation and agency. If I'm right, understanding what seems like a fundamental, scientific or metaphysically deep feature of the world requires thinking about us—about the features and activities of agents. This is to take a broadly pragmatist or idealist stance, in the sense that we understand a putatively fundamental, scientific or metaphysical feature of the world by thinking about its relevance for our lives. Accounts that relate agency to such features can be surprising and illuminating. But they don't make the features less real or less than fully objective.

If this approach works in the case of causation, where else might it apply? I'd like to end by suggesting a broad program for applying this form of pragmatism to other areas of science. Firstly, it may help us make sense of scientific laws. We employ standards to pick out laws. Fundamental physical laws, for example, are often taken to be universal, exceptionless regularities that are both simple in form and general in scope. Lewis (1994), for example, explicitly defends a 'Best Systems Account' of this form. But why should we expect nature to operate by regularities that obey these standards? Under this pragmatist approach, we can justify these standards by appealing to agency. It's because we are reasoners with limited cognitive capacities that we reason to laws that are simple and general. This doesn't make scientific laws mind-dependent. But it does mean we can't understand features of laws without thinking about agents.

Similar points hold for making sense of scientific explanation and theory choice. Why do we seek causal explanations? A pragmatist approach suggests that causal explanations matter to us because we are interested in deliberation and control. Causal explanations do not aim at uncovering the fundamental structure of the world. Why do we seek explanations based in laws? A pragmatist approach might suggest that nomic explanations are useful because they help us *unify* our knowledge of the world, given that laws are appropriately general and simple. This approach could be used to defend unificationist accounts of explanation (Kitcher 1981), in combination with system-based accounts of laws (Lewis 1994).

This kind of pragmatist approach also suggests that standards of coherence will be important for making sense of scientific enquiry—particularly in theory choice. We might expect to find these standards at work when we require a theory's truth not to undermine our reasons for holding it—as occurs in 'Boltzmann Brain' scenarios in statistical mechanics or versions of quantum mechanics. The importance of coherence requirements may also lead us to reject *a priori* constraints on scientific theories, such as requirements to treat spacetime as fundamental—as found in some defences of 'primitive ontology' in quantum mechanics. Coherence suggests that our standard for science should evolve with science.

Altogether, this pragmatist approach aims to make sense of science by thinking about its human side—about the ways in which our needs and interest shape the kind of science we produce. The deliberative account of causation is one example of where such an approach might lead. It is because we are deliberators, reasonably able to decide on one thing in order to achieve another, that we view the world in causal terms. And so we can't make sense of causation independently of our concerns and limits as agents.

**References**

Ahmed, Arif. 2005. Evidential Decision Theory and Medical Newcomb Problems. *British Journal for the Philosophy of Science* 56: 191–8.

Ahmed, Arif. 2014. *Evidence, Decision and Causality*. Cambridge: Cambridge University Press.

Albert, David Z. 1994. Quantum Mechanics and the Approach to Thermodynamic Equilibrium. *British Journal for the Philosophy of Science* 45: 669–677.

———. 2000. *Time and Chance*. Cambridge, Mass.: Harvard University Press.

———. 2014. The Sharpness of the Distinction between Past and Future. In *Asymmetries of Chance and Time*. ed. Alistair Wilson. Oxford: Oxford University Press.

———. 2015 *After Physics*. Cambridge, Mass.: Harvard University Press.

Anscombe, G. E. M. 1957. *Intention*. Cambridge, Mass.: Harvard University Press.

———. 1969. Causality and Extensionality. *The Journal of Philosophy* 66(6): 152–159.

Aristotle. *Nicomachean Ethics*. 1925. In *The Nicomachean Ethics*. trans. David Ross. Oxford: Oxford University Press.

Armstrong, David. 1983. *What Is a Law of Nature?* Cambridge: Cambridge University Press.

Beebee, Helen, Hitchcock, Chris and Price, Huw. eds. 2000. *Making a Difference: Essays in Honour of Peter Menzies*. Oxford: Oxford University Press.

Beiser, Frederick. 2000. The Enlightenment and Idealism. In *The Cambridge Companion to German Idealism*. ed. Karl Ameriks. Cambridge: Cambridge University Press.

Bennett. Jonathan .1984. Counterfactuals and Temporal Direction. *The Philosophical Review*. 93(1): 57–91.

———. 1988. *Events and Their Names*. Oxford: Clarendon Press.

Bok, Hilary. 1998. *Freedom and Responsibility*. Princeton: Princeton University Press.

Bratman, Michael. 1984. Two Faces of Intention. *The Philosophical Review* 93(3): 375–405.

———. 1991. Cognitivism about Practical Reason. *Ethics* 102(1): 117–128.

Callender, Craig. 2004. Measures, Explanation and the Past: Should 'Special' Initial Conditions Be Explained? *British Journal for the Philosophy of Science* 55: 195–217.

Carroll, John. 1994. *Laws of Nature*. Cambridge: Cambridge University Press.

———. 2010. *From Eternity to Here*. New York: Dutton.

Cartwright , Nancy. 1979. Causal Laws and Effective Strategies. *Noûs* 13(4): 419–437.

———. 2003. Two Theorems on Invariance and Causality. *Philosophy of Science* 70: 203–24.

Coffman, E. J. and Warfield, Ted. A. 2005. Deliberation and Metaphysical Freedom. *Midwest Studies in Philosophy* 29: 25–44.

Collingwood, R. G. 1940. *An Essay on Metaphysics.* Oxford: Clarendon Press.

Cox, J. W. Roxbee. 1963. Can I Know Beforehand What I am Going to Decide? *Philosophical Review* 72(1): 88–92.

Davidson, Donald. 1980. *Essays on Actions and Events.* Oxford: Clarendon Press.

Dennett, Daniel. 1984. *Elbow Room.* Cambridge, Mass.: MIT Press.

Dowe, Philip. 1992a. Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory. *Philosophy of Science* 59: 195–216.

———. 1992b. Process Causality and Asymmetry. *Erkenntnis* 37(2): 179–196.

Dummett, Michael. 1954. Can an Effect Precede its Cause. *Proceedings of the Aristotelian Society* 28(Supplement): 27–44.

———. 1964. Bringing about the Past. *The Philosophical Review* 73(3): 338–359.

Earman, John. 1974. An attempt to Add and Little Direction to 'The Problem of the Direction of Time.' *Philosophy of Science.* 41: 15–47.

———. 1976. Causation: A Matter of Life and Death. *The Journal of Philosophy* 73(1): 5–25.

———. 1986. *A Primer on Determinism.* Dordrecht: Reidel.

———. 2006. The "Past Hypothesis": Not Even False. *Studies in History and Philosophy of Modern Physics* 37: 399–430.

Eells, Ellery. 1981. Causality, Utility, and Decision. *Synthese* 48(2): 295–329.

———. 1982. *Rational Decision and Causality.* Cambridge: Cambridge University Press.

———. 1984. Metatickles and the Dynamics of Deliberation. *Theory and Decision* 17(1): 71–95.

Elga, Adam. 2001. Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science* 68(3): S313-S324.

Fernandes, Alison. Forthcoming$_1$. Time, Flies, and Why We Can't Influence the Past. In *Time's Arrows and the Probability Structure of the World* ed. Barry Loewer, Brad Weslake and Eric Winsberg. Cambridge, Mass.: Harvard University Press.

———. Forthcoming$_2$. Varieties of Epistemic Freedom. *Australasian Journal of Philosophy.*

Feynman, Richard. 1965. *The Character of Physical Law.* Cambridge, Mass.: MIT Press.

Fichte, Johann Gottlieb [1796/99] 1992. *Wissenschaftslehre nova methodo.* In *Foundations of Transcendental Philosophy (Wissenschaftslehre) nova method* ed. and trans. Daniel Brezeale. Ithaca: Cornell University Press.

———. [1797/98] 1994. *Versuch einer neuen Darstellung der Wissenschaftslehre.* In *An Attempt at a New Presentation of the Wissenschaftslehre* ed. and trans. Daniel Breazeale. Indianapolis: Hackett.

Field, Hartry. 2003. Causation in a Physical World. In *The Oxford Handbook of Metaphysics.* ed. Michael J. Loux and Dean W. Zimmerman. Oxford: Oxford University Press. 435–460.

Frankfurt, Harry. 1971. Freedom of the Will and the Concept of a Person. *Journal of Philosophy* 68(1): 5–20.

Frisch, Mathias. 2007. Causation, Counterfactuals, and the Past-Hypothesis. In *Causation, Physics, and the Constitution of Reality.* ed. Huw Price and Richard Corry. Oxford*:* Oxford University Press. 351–396.

———. 2010. Does a Low-Entropy Constraint Prevent Us from Influencing the Past? In *Time, Chance and Reduction.* ed. Gerhard Ernst and Andreas Hüttemann. Cambridge: Cambridge University Press. 13–33.

Gasking, Douglas. 1955. Causation and Recipes. *Mind* 64: 479–487.

Gilboa, I. 1999. Can Free Choice be Known? In *The Logic of Strategy.* ed. Cristina Bicchieri, Richard Jeffrey and Brian Skyrms. Oxford: Oxford University Press.

Ginet, Carl. 1962. Can The Will be Caused? *The Philosophical Review* 71(1): 49–55.

Goldman, Alvin. 1970. *A Theory of Human Action.* Englewood Cliffs, New Jersey: Prentice-Hall.

Greene, Brian. 2004. *The Fabric of the Cosmos.* New York: Alfred A. Knopf.

Grice, H. P. 1971. Intention and Uncertainty. *Proceedings of the British Academy* 57: 263–79.

Grünbaum, Adolf. 1963. *Philosophical Problems of Space and Time.* New York: Knopf.

Hall, Ned. 2004. Two Concepts of Causation. In *Causation and Counterfactuals.* ed. John Collins, Ned Hall and Laurie Paul. Cambridge, Mass: MIT Press. 225–76.

———. 2007. Structural Equations and Causation. *Philosophical Studies* 132(1): 109–136.

Hampshire, Stuart and Hart, H. L. A. 1958. Decision, Intention and Certainty. *Mind* 67(265): 1–12.

Harman, Gilbert. 1976. Practical Reasoning. *The Review of Metaphysics* 29(3): 431–463.

Hitchcock, Christopher. 1996. Causal Decision Theory and Decision-theoretic Causality. *Noûs* 30: 508–26.

———. 2007. What Russell Got Right. In *Causation, Physics, and the Constitution of Reality*. ed. Huw Price and Richard Corry. Oxford*:* Oxford University Press. 45–65.

Holton, Richard. 2006. The Act of Choice. *Philosophers' Imprint* 6(3): 1–15.

Horwich, Paul. 1987. *Asymmetries in Time: Problems in the Philosophy of Science*. Cambridge, Mass.: MIT Press.

Hume, David. [1738] 2000. *A Treatise of Human Nature.* eds. David Fate Norton, Mary J. Norton. Oxford; New York: Oxford University Press.

Hume, David. [1748] 1999. *An Enquiry Concerning Human Understanding.* ed. Tom L. Beauchamp. Oxford; New York: Oxford University Press.

Ismael, Jenann. 2007. Freedom, Compulsion and Causation. *Psyche* 13(1): 1–11.

———. 2012. Decision and the Open Future. In *The Future of the Philosophy of Time*. ed. A. Bardon. London: Routledge. 149–68.

———. 2016. How Do Causes Depend on Us? The many faces of perspectivalism. *Synthese* 193(1): 245–267.

Jeffrey, Richard C. 1977. A Note on the Kinematics of Preference. *Erkenntnis* 11(1): 135–141.

Jones, David. 1968. Deliberation and Determinism. *Southern Journal of Philosophy* 6: 255–64.

Joyce, James M. 2002. Levi on Causal Decision Theory and the Possibility of Predicting One's Own Actions. *Philosophical Studies* 110(1): 69–102.

———. 2007. Are Newcomb Problems Really Decisions? *Synthese* 156(3): 537–62.

Kant, Immanuel. [1781/1787] 1996. *Critique of Pure Reason*. trans. Werner S. Pluhar. Indianapolis: Hackett.

Kapitan, Tomis. 1984. Can God Make up his Mind? *International Journal for Philosophy of Religion* 15: 37–48.

———. 1986. Deliberation and the Presumption of Open Alternatives. *The Philosophical Quarterly* 36(143): 230–51.

———. 1989. Doxastic Freedom: A Compatibilist Alternative. *American Philosophical Quarterly* 26: 31–41.

———. 1990. Action, Uncertainty, and Divine Impotence. *Analysis* 50(2): 127–133.

Kitcher, Philip. 1981. Explanatory Unification. *Philosophy of Science* 48: 507–531.

Korsgaard, Christine. 1996. *The Sources of Normativity.* Cambridge: Cambridge University Press.

Kutach, Douglas. 2002. The Entropy Theory of Counterfactuals. *Philosophy of Science* 69(1): 82–104.

———. 2007. The Physical Foundations of Causation. In *Causation, Physics, and the Constitution of Reality*. ed. Huw Price and Richard Corry. Oxford*:* Oxford University Press. 327–350.

Langton, Rae. 2004. Intention As Faith. *Royal Institute of Philosophy Supplement* 55: 243–258.

Lavin, Douglas. 2013. Must There Be Basic Action? *Noûs* 47(2): 273–30.

Leeds, Stephen. 2003. Foundations of Statistical Mechanics—Two Approaches. *Philosophy of Science* 70(1): 126–144.

Levi, Isaac. 1980. *The Enterprise of Knowledge*. Cambridge, Mass.: MIT Press.

———. 1986. *Hard choices*. Cambridge: Cambridge University Press.

———. 1997. *The Covenant of Reason*. Cambridge: Cambridge University Press.

———. 2000. Review of James M. Joyce: The Foundations of Causal Decision Theory. *Journal of Philosophy* 97(7): 387–402.

Lewis, David. 1973a. Causation. *Journal of Philosophy* 70: 556–67. (reprinted with postscripts in Lewis 1986.)

———. 1973b. *Counterfactuals*. Oxford: Blackwell.

———. 1976. The Paradoxes of Time Travel. *American Philosophical Quarterly* 13(2): 145–152.

———. 1979a. Counterfactual Dependence and Time's Arrow. *Nous* 13: 455–76. (reprinted with postscripts in Lewis 1986.)

———. 1979b. Are We Free to Break the Laws? *Theoria: A Swedish Journal of Philosophy* 47: 113–121.

———. 1983. New Work for a Theory of Universals *Australasian Journal of Philosophy* 61: 343–377.

———. 1986. *Philosophical Papers: Volume II*. New York: Oxford University Press.

———. 1986a. A Subjectivist Guide to Objective Chance. In Lewis 1986. 114–132.

———. 1986b. Causal Explanation. In Lewis 1986. 214–40.

———. 1994. Humean Supervenience Debugged. *Mind* 103(412): 473–390.

———. 2000. Causation as Influence. *Journal of Philosophy* 97(4): 182–97.

Loewer, Barry. 2001. Determinism and Chance. *Studies in the History of Modern Physics* 32(4): 609–620.

———. 2007. Counterfactuals and the Second Law. In *Causation, Physics, and the Constitution of Reality*. ed. Huw Price and Richard Corry. Oxford: Oxford University Press. 293–326.

———. 2012. Two accounts of Laws and Time. *Philosophical Studies* 160(1): 115–137.

Maudlin, Tim. 2002. *Quantum Non-Locality and Relativity*. 2nd ed. Malden, Mass.: Blackwell.

———. 2007. *The Metaphysics within Physics*. Oxford: Oxford University Press.

McCann, Hugh. 1975. Is Raising One's Arm a Basic Action? *Journal of Philosophy* 69(9): 235–249.

Mellor, D. Hugh. 1988. On Raising the Chances of Effects. In *Probability and Causality*. ed. James H. Fetzer. Dordrecht: Reidel. 229–39.

———. 1995. *The Facts of Causation*. London: Routledge Press.

Menzies, Peter and Price, Huw. 1993. Causation as a Secondary Quality. *British Journal for the Philosophy of Science* 44: 187–203.

Moran, Richard. 2001. *Authority and Estrangement*. Princeton: Princeton University Press.

Murray, A. H. and Burchfield, R. W. eds. 1933. *The Oxford English Dictionary*. Oxford: Clarendon Press.

Nozick, Robert. 1969. Newcomb's Problem and Two Principles of Choice. In *Essays in Honour of Carl G. Hempel*. ed. N. Rescher. Dordrecht: Reidel. 114–146.

O'Brien, Lucy and Soteriou, Matthew. eds. 2009. *Mental Actions*. Oxford: Oxford University Press.

O'Shaughnessy, Brian. 1980. *The Will: A Dual Aspect Theory*. Cambridge: Cambridge University Press.

Owens, David. 2008. Deliberation and the First Person. In *Self-Knowledge*. ed. Anthony E. Hatzimoysis. Oxford: Oxford University Press.

———. 2009. Freedom and Practical Judgement. In *Mental Actions*. ed. Lucy O'Brien and Matthew Soteriou. Oxford: Oxford University Press.

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

Pearl, Judea and Verma, T. S. 1991. A Theory of Inferred Causation. In *Principles of Knowledge Representation and Reasoning* ed. J. A. Allen, R. Fikes and E . Sandewall. San Mateo, Cal.: Morgan Kaufmann. 441–52.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Pears, David. 1968. Predicting and Deciding. In *Studies in the Philosophy of Thought and Action*. ed. P. F Strawson. Oxford: Oxford University Press.

Penrose, Roger. 2005. *The Road to Reality*. New York: Alfred A. Knopf.

Pereboom, Derek. 1995. Determinism Al Dente *Nous* 29: 32–33.

Perry, David. 1965. Prediction, Explanation and Freedom. *Monist* 49(2): 234–47.

Perry, John. 1979. The Problem of the Essential Indexical. *Noûs* 13: 3–21.

Pettit, Philip. 1991. Realism and Response-Dependence. *Mind* 100: 587–626.

Price, Huw and Weslake, Brad. 2009. The Time-Asymmetry of Causation. In *The Oxford Handbook of Causation*. ed. Helen Beebee, Christopher Hitchcock and Peter Menzies. Oxford*:* Oxford University Press.

Price, Huw. 1986. Against Causal Decision Theory. *Synthese* 67(2): 195–212.

———. 1988. *Facts and the Function of Truth*. Oxford: Blackwell.

———. 1991. Agency and Probabilistic Causality. *British Journal for the Philosophy of Science* 42: 157–76.

———. 1992. Agency and Causal Asymmetry *Mind* 101: 501–520.

———. 1993. The Direction of Causation: Ramsey's Ultimate Contingency. In *PSA 1992 Volume 2* ed. David Hull, Micky Forbes and Kathleen Okruhlik. East Lansing, Michigan: Philosophy of Science Association.

———. 1996. T*ime's Arrow and Archimedes' Point: New Directions for the Physics of Time*. Oxford: Oxford University Press.

———. 2002. Boltzmann's Time Bomb. *British Journal for the Philosophy of Science* 53: 83–119.

———. 2007. Causal Perspectivalism. In *Causation, Physics, and the Constitution of Reality*. ed. Huw Price and Richard Corry. Oxford*:* Oxford University Press. 250–292.

———. 2012. Causation, Chance and the Rational Significance of Supernatural Evidence. *Philosophical Review* 121(4): 483–538.

———. Forthcoming. Causation, Intervention and Agency—Woodward on Menzies and Price. In *Making a Difference*. ed. Helen Beebee, Chris Hitchcock and Huw Price. Oxford: Oxford University Press.

Prior, Arthur N. 1968. *Papers on Time and Tense*. Oxford: Oxford University Press.

Rabinowicz, Wlodek. 2002. Does Practical Deliberation Crowd Out Self-Prediction? *Erkenntnis* 57(1): 91–122.

Ramsey, Frank P. [1929] 1978. General Propositions and Causality. In *Foundations: Essays in Philosophy, Logic, Mathematics and Economics* ed. D. H. Mellor. London: Routledge & Kegan Paul. 133–51.

Reichenbach, Hans. [1956] 1991. *The Direction of Time*. Mineola: Dover Publications.

Roessler, Johannes. 2013. The Epistemic Role of Intentions. *Proceedings of the Aristotelian Society* 113: 39–56.

Russell, Bertrand. [1910] 2009. The Elements of Ethics. In *Philosophical Essays*. London, New York: Routledge. 3–51.

———. 1912–13. On the Notion of Cause. *Proceedings of the Aristotelian Society, New Series* 13: 1–26.

———. [1914] 1996. *Our Knowledge of the External World*. London, New York: Routledge.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Schick, Frederic. 1979. Self-knowledge, Uncertainty, and Choice. *British Journal for the Philosophy of Science* 30: 235–252.

Schroeder, Mark. 2011. Ought, Agents, and Actions. *Philosophical Review* 120(1): 1–41.

Searle, John R. 2001. *Rationality in Action*. Cambridge, Mass.: MIT Press

Setiya, Kieran. 2008. Practical Knowledge. *Ethics* 118: 388–409.

Shackle, G. L. S. 1958. *Time in Economics*. Amsterdam: North Holland Pub. Co.

Silverberg, Robert. 1975. *The Stochastic Man*. New York: Harper & Row.

Sklar, Lawrence. 1993. *Physics and Chance*. Cambridge: Cambridge University Press.

Skyrms, Brian. 1980. *Causal Necessity*. New Haven: Yale University Press.

———. 1990. *The Dynamics of Rational Deliberation*. Cambridge, Mass.: Harvard University Press.

Smart, J. J. C. 1967. Time. *Encyclopedia of Philosophy. Vol. 8*. New York: Macmillan.

Smith, Matthew. 2010. Practical Imagination and its Limits. *Philosophers' Imprint* 10(3): 1–20.

Smith, Nicholas J. J. 2005. Why Would Time Travellers Try to Kill their Younger Selves? *Monist* 88(3): 388–95.

Sorensen, Roy A. 1984 Uncaused Decisions and Pre-Decisional Blindspots. *Philosophical Studies* 45(1): 51–56.

Spinoza, Benedict de. [1677] 1996. *Ethica, Ordine Feometrico Demonstrate*. In *Ethics*. ed. and trans. Edwin Curley. London: Penguin.

Spirtes, Peter, Glymour, Clark and Scheines, Richard. 1993. *Causation, Prediction and Search*. New York: Springer-Verlag.

Spohn, Wolfgang. 1977. Where Luce and Krantz do Really Generalize Savage's Decision Model. *Erkenntnis* 11: 113–134.

———. 2007. The Core of Free Will. In *Thinking about Causes*. ed. P. K Machamer and G. Wolters. Pittsburgh: Pittsburgh University Press.

Stocker, Michael. 1968. Knowledge, Causation and Decision. *Noûs* 2(1): 65–73.

Strawson, Peter F. 1974. Freedom and Resentment. In *Freedom and Resentment and Other Essays*. London: Methuen.

Strevens, Michael. 2007. Review of Woodward, 'Making Things Happen'. *Philosophy and Phenomenological Research* 74(1): 233–249.

———. 2008. *Depth*. Cambridge, Mass.: Harvard University Press.

Taylor, Richard. 1964. Deliberation and Foreknowledge. *American Philosophical Quarterly* 1(1): 73–80.

Tooley, Michael. 1977. The Nature of Laws. *Canadian Journal of Philosophy* 7: 667–698.

Ullmann-Margalit, Edna and Morgenbesser, Sidney. 1977. Picking and Choosing. *Social Research* 44(4): 757–785.

van Inwagen, Peter. 1983. *An Essay on Free Will*. Oxford: Oxford University Press.

———. 2008. *Metaphysics*. Boulder, Colo.: Westview Press.

Velleman, J. David. 1985. Practical Reflection. *The Philosophical Review* 94(1): 33–61.

———. 1989a. *Practical Reflection*. Princeton: Princeton University Press.

———. 1989b. Epistemic Freedom. *Pacific Philosophical Quarterly* 70: 73–97.

———. 2000. *The Possibility of Practical Reason*. Oxford: Oxford University Press.

von Wright, Georg. 1971. *Explanation and Understanding*. Ithaca, New York: Cornell University Press.

Weslake, Brad. 2006. Review of Making Things Happen. *Australasian Journal of Philosophy* 84(1): 136–140

———. 2014. Statistical-mechanical Imperialism. In *Asymmetries of Chance and Time*. ed. Alistair Wilson. Oxford: Oxford University Press.

Williams, Bernard. 2002. *Truth and Truthfulness*. Princeton: Princeton University Press.

Winsberg. Eric. 2004. Can Conditioning on the Past Hypothesis Militate Against the Reversibility Objections? *Philosophy of Science* 71 (4): 489–504.

Wittgenstein, Ludwig. [1921] 1961. *Logisch-Philosophische Abhandlung*. In *Tractatus Logico-Philosophicus* trans. D. F. Pears and B. F. McGuinness. New York: Humanities Press.

Wood, Allen. 1984. Kant's Compatibilism. In *Self and Nature in Kant's Philosophy*. ed. Allen Wood. Ithaca and London: Cornell University Press. 73–101.

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

———. 2007. Causation with a Human Face. In *Causation, Physics, and the Constitution of Reality*. ed. Huw Price and Richard Corry. Oxford*:* Oxford University Press. 66–105.

———. 2013. Causation and Manipulability. *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition) ed. Edward N. Zalta. URL = <http://plato.stanford.edu/archives/win2013/entries/causation-mani>.