Preprint—Forthcoming in *Philosophy and Phenomenological Research*

A Deliberative Approach to Causation

Alison Fernandes

Abstract

Fundamental physics makes no clear use of causal notions; it uses laws that operate in relevant respects in both temporal directions and that relate whole systems across times. But by relating causation to evidence, we can explain how causation fits in to a physical picture of the world and explain its temporal asymmetry. This paper takes up a deliberative approach to causation, according to which causal relations correspond to the evidential relations we need 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. This correspondence explains why causation matters: knowledge of causal structure helps us make decisions that are evidence of outcomes we seek. The account also explains why we can control the future and not the past, and why causes come *before* their effects. When agents properly deliberate, their decisions can never count as evidence for any outcomes they may seek in the past. From this it follows that causal relations don't run backwards. This deliberative asymmetry is itself traced back to asymmetries of evidence and entropy, providing a new way of deriving causal asymmetry from temporally symmetric laws.

1. Introduction

Causation plays an essential role in our scientific and everyday understanding of the world. We discover causal relations using experiments in labs, studies and everyday life—such as when a chemist tests whether adding excess acid causes the wrong product to precipitate, or

when statistical data are used to determine if a new drug decreases cholesterol absorption, or when you experiment with having less coffee to see if it improves your sleep. We also construct theories that describe and explain causal relations. Biologists map the causal relations involved in photosynthesis, physicists explain why absorbing a photon causes an electron to jump to a higher energy level, and you might hypothesise about how too much water is affecting your plant's health. Talk of causation is ubiquitous.

So it might come as a surprise 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 theories, causal talk doesn't appear. Fundamental physical theories don't identify particular states as causes, and others as effects. Instead they use dynamical equations that relate global states of affairs—no mention of causes. Furthermore, when we look more closely at fundamental physical laws, they seem to have precisely the wrong features to deliver causal relations. Firstly, the laws are not *local* in the way causes are. It might seem that if a billiard ball *A* collides with stationary ball *B*, causing it to move off, the local state of *A* moving *necessitates B*'s motion, given the fundamental laws. But, as Russell famously argued (1912−13), fundamental laws only determine states of affairs at other times given information about the *entire* global state of affairs. *B*'s motion also depends on the fact every other billiard ball in the system fails to collide with it. Laws don't deliver local necessitation. Secondly, fundamental laws don't reflect the temporal asymmetry of causation. It might seem that fundamental laws only necessitate states towards the future: a state involving *A*'s earlier motion necessitates *B*'s in a way that a state involving *B*'s later motion doesn't necessitate *A*'s. But fundamental physical laws, in relevant respects, work equally well in *both*

temporal directions.[1] Altogether, given our intuitive view of causation, it's unclear how causation fits into the picture of the world presented by fundamental physics. Even if we don't follow Russell in trying to eliminate causal talk altogether, we need to look somewhere other than the predictions and derivations of fundamental physics to understand causation. I'll take this to be Russell's challenge.

A promising strategy to respond to Russell's challenge has been to tie causation to counterfactuals and control. According to the general form of counterfactual accounts, effects depend counterfactually on their causes, such that by manipulating a cause we can manipulate its effects. If causation has this kind of *practical* import, we can make sense of why we need causal relations, in addition to the laws of fundamental physics. Fundamental physical laws capture the most basic and universal exceptionless regularities—those that hold across whole patterns of events. Causal structure captures how manipulating one local state can be a means of manipulating another—relations that higher-level sciences are often more interested in. If causation is tied to manipulation, we can understand why causal relations are needed, even if they don't feature explicitly in fundamental physics. Moreover, given how we rely on causal reasoning in decision-making, making sense of the practical import of causation is an important independent desideratum on accounts of causation.

The account I'll go on to defend shares features in common with a broad class of counterfactual accounts that take causation to have a practical import. But it aims to do

[1] For further discussion of Russell's arguments, see Earman (1976), Field (2003) and Eagle (2007). While Russell takes the fundamental laws to be Newtonian, similar arguments hold for better candidates. For example, even if the laws are temporally asymmetric in minimal ways (such as to allow asymmetry in neutral K-meson decay), this doesn't straightforwardly explain the pervasive macroscopic asymmetry of causation. If we exclude faster-than-light influence (although not actually a consequence of relativity) the information required to determine an effect is still too large to pick out individual causes.

better in one crucial respect. It is not enough to simply claim that the relations picked out as causal are relevant to our practical lives, or merely stipulate that the relations are ones of control or counterfactual dependence. One has to actually *show* that the relations picked out as causal or counterfactual *are* those that matter to our practical lives. There are many relations we might have picked out as causal, and many ways we might have evaluated the relevant counterfactuals. Why are these the right ones? How do they matter to us? Can we show by some independent means that we *should* pick out these relations as causal?

This criterion is typically not explicitly appealed to in giving accounts of causation (although it plays a role). More often defenders are interested in whether a given analysis delivers our intuitive verdicts on core cases, or whether the relation is used in scientific practice. But it is an important criterion nevertheless in showing that an account picks out the relation it needs to. Analogous criteria appear in other areas of philosophy, where accounts must explain how a given property or relation fits into our picture of the world, in a way that explains why it should matter to us as it does. Accounts from meta-ethics, for example, should make sense of why ethical properties matter to our motivations; accounts of colour should makes sense of how color properties are perceptible. It is not enough to simply claim they are. Sometime the criterion is appealed to in order to attack other accounts: it is a point against a meta-ethical account, for example, if an ideally rational agent could accept an ethical proposition, but fail to be moved by it. Here I'm appealing to the criterion as a positive standard. It is not only a problem when the relations picked out by an account as causal manifestly fail to matter to our practical lives. Defenders must actually *show* how the relations matter. Without this, an account does not adequately show how causation fits into our picture of the world.

However, it turns out that popular counterfactual accounts of causation don't explain how the relations they pick out as causal matter to us. First, consider *reductive* counterfactual accounts—those that attempt to reduce causal relations to non-causal relations via counterfactuals. David Lewis' account (1973, 1979) exemplifies this approach. According to Lewis, a causal relation obtains between two events just in case there is a chain of counterfactual dependencies between them. One event **A** depends counterfactually on another independent event **B**, just in case were **A** to occur **B** would occur and were **A** not to occur **B** would not occur. Lewis introduces a comparative similarity ordering between possible worlds to evaluate these counterfactuals. Roughly, a counterfactual is true at a world if the closest world where the antecedent is true is also one where the consequent is true. Lewis takes the similarity ordering between worlds to be a metaphysical primitive. But he introduces standards by which similarity is evaluated, such as 'avoid big, widespread, diverse violations of law', and 'it is of little or no importance to secure approximate similarity of particular fact' (1979, p. 472). He needs these standards in order to complete his reductive account.

But there's a problem. As Bennett (1984) and others note, Lewis gives us no *justification* for these standards. Why care about perfect match far more than imperfect match? Why allow for violations of the laws, if we can't actually violate them? These criteria look arbitrary and leave it mysterious why causation, understood in these terms, should matter to our lives. Lewis takes his criteria to deliver the 'standard' resolution of the vagueness of counterfactuals (1979, p. 457). He is also explicit that these criteria have been reverse-engineered to deliver our intuitive causal judgments (ibid., pp. 466−7). But he does not explain why the criteria are appropriate, or why our intuitive judgments are worth following.

If someone uses different criteria and claims he would float, were he to jump out a window, we need to say more than that he is evaluating counterfactuals in a non-standard way.[2] Even if Lewis' account gets the extension of the concept CAUSATION right, it doesn't explain the practical import of causation.

Other reductive accounts inherit this problem, including the statistical-mechanical accounts of David Albert (2000, 2015) and Barry Loewer (2007). These counterfactual accounts use results from statistical mechanics to explain causal asymmetry—avoiding problems with Lewis' account raised by Elga (2001). However, they still do not justify the terms of their reductions. Loewer evaluates decision counterfactuals by holding the present macroscopic state of the world fixed and introducing changes to the microstate within the brain to model decisions. This means that changes to the microstate of the brain cannot be correlated with changes in the surrounding environment. But why should we rule out such correlations? Albert uses a different method for evaluating counterfactuals, one that doesn't involve holding the present macrostate fixed. But he introduces an unanalysed 'fiction of agency' and employs similar standards, such as keeping the initial macrostate of the universe fixed. These standards need to be justified if we're to understand why causation, understood in these terms, should matter to us—and so fully address Russell's challenge.[3]

What about non-reductive accounts? Judea Pearl (2000) and James Woodward (2003) defend interventionist accounts that relate causal relations to one another using 'interventionist

[2] This is a particular problem for Lewis, who uses counterfactuals to formulate rational decision theory (1981).
[3] The nearest Albert comes to justifying the reduction is to claim that counterfactuals are evaluated using our 'normal procedures of inference' (2000, p. 129). But then the account faces the same problems as the deliberative approach regarding why evidential correlations to the past can't be exploited (section 4). See Frisch (2010) for more on this objection.

counterfactuals' (rather than reduce causal relations to non-causal relations). These accounts are scientifically motivated and provide valuable analyses of causal concepts. They're also advertised as doing well at explaining the practical import of causation. Here's Woodward (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.

A later paper by Woodward is titled 'A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment)' (2014). Similar remarks can be found in Pearl (2000, p. 337). But these accounts don't actually adequately explain how causation matters to us. Here's why. According to the general form of interventionist accounts, **A** is causally related to **B** if and only if **A** would remain correlated with **B**, were **A** to be intervened on by a suitable causal process (where **A** and **B** are independent variables). A suitable causal process is one that one that breaks causal relations upstream, but leaves other relevant relations intact. Woodward and Pearl give precise technical (and competing) specifications for what counts as an intervention—and so what it takes for a causal relation to hold. But they don't show that human actions or decisions ever satisfy the conditions on interventions, except by stipulation (Pearl 2000, pp. 108−9). And so they don't show that interventionist counterfactuals are ones we can make

use of. Nor do they explain why it should matter to us that our actions *do* satisfy these conditions—why interventionist counterfactuals are ones we should care about. Woodward sometimes claims that the rationale for his requirements is 'commonsensical' (2014, p. 706). But the requirements on interventions are controversial, and require detailed specification. They are not intuitive. Price and Weslake (2009) raise similar concerns.

At some points, Woodward claims interventionist relations are useful because they are 'potentially exploitable for purposes of manipulation' (2003, p. 7) and 'control' (2014, p. 696). But he doesn't give an independent specification of manipulation or control. He simply assumes these are to be explicated in interventionist terms. But if one is worried that these interventionist counterfactuals are not latching onto useful relations, it is no comfort to be told that the relations they latch onto are useful *because* they are interventionist counterfactuals. Their usefulness of these counterfactuals is precisely what's in question. Appealing to manipulation and control *characterised in interventionist terms* is too small a circle to be explanatory. While Woodward does explain why certain features of causal relations should matter to us, like stability across changes in background conditions (2000, ch. 6, 2014, p. 704), he doesn't justify why interventionist counterfactuals should matter to us in general.[4] Interventionist accounts, on their own, don't justify why we should pick out certain relations as causal.

One could begin with an interventionist or reductive counterfactual account and attempt to show that it picks out a useful relation. In this paper, I take a different approach. Rather than

[4] Woodward claims that if we don't pick out interventionist causal relations, 'it's entirely possible that when we perform [a] manipulation [on X in a new setting] X and Y will no longer be correlated' (2014, p. 709). But it's always possible that causal relations break down in new settings. Nor does the claim justify why Woodward's interventionism is to be preferred over any other account that requires causal relations to be stable.

begin with an account of causation and try to explain why *that* relation will be useful to agents, I'll begin by taking the practical import of causation to be central, and characterize causation in terms of *how* it is useful. This is to take a broadly pragmatist approach to causation. According to the 'deliberative approach', causal relations correspond to the evidential relations agents use when they decide on one thing in order to achieve another. Say Tamsin is deliberating about whether to take her umbrella in order to stay dry. She needs to know whether her decision to take her umbrella (in order to stay dry) is generally adequate grounds for thinking she'll stay dry. If it is, she'll have made a good decision. According to the deliberative approach, a causal relation obtains between her taking her umbrella and staying dry, just in case this evidential relations obtains (and Tamsin is 'properly deliberating'). We care about causal structure because it directs us to decisions that are evidence of outcomes we seek. Causation matters to us because we need it for decision-making.

This deliberative approach has some unusual features. Firstly, it takes causation to be foremost tied to *inference*. It tracks when one state of affairs is good evidence of another, rather than when one state necessitates or produces another. Counterfactual accounts already move us away from a picture where causes are needed to push and pull things around at the fundamental level. The deliberative account goes further, and relates causation to evidential (rather than deep physical or metaphysical) structure. Secondly, the account is relatively hands off about the metaphysics of causation. While I'll defend a particular biconditional relating causation to deliberation, this biconditional is compatible with different accounts of what causal relations are. It could be used to justify a first-order account of causation or even reduce causal relations to evidential relations. I won't argue

against these metaphysical projects here. What I will show is that the deliberative account does important explanatory work independently of settling the metaphysical nature of causation—in this sense, it offers an independent account. By relating causation to evidence, the deliberative account explains why causation matters, how it relates to fundamental laws, and how it fits into a scientific picture of the world. Such an account is worth exploring even if one thinks causation is primitive or otherwise reducible. Thirdly, the approach highlights an important middle-ground between first-order realist accounts of causation (including interventionist and reductive accounts) and standard agent-based accounts. First-order accounts don't typically aim to explain why relations matter to us. Agent-based accounts have typically been used to defend forms of subjectivism or anti-realism. The deliberative approach uses agential standards to pick out objective relations, explain their temporal features, and reconcile them with fundamental physics. My aim here isn't to argue that first-order or agent-based accounts are false; I offer an alternative way of doing scientifically informed philosophy.

The paper proceeds as follows. Section 2 lays out the deliberative approach in detail. Section 3 considers how the approach differs from other agent-based accounts and how it delivers objective causal relations. Section 4 uses features of deliberation and evidence to explain why causes come before their effects and why we control the future and not the past.

2. The Deliberative Approach

2.1. From Evidence to Causation

According to the deliberative approach, causal relations correspond to the evidential relations we use when we decide on one thing in order to achieve another. The following

evidential biconditional characterises this correspondence, where **A**, deciding on **A**, and **B** are independent states of affairs, finely individuated:[5]

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

The evidential biconditional implies that for two states of affairs to be causally related, an agent's deciding on one must be good evidence for the other. Say Tamsin is properly deliberating over whether to take her umbrella. According to the biconditional, her taking her umbrella is a *cause* of her not getting wet just in case her deciding on taking her umbrella (for the sake of not getting wet) is good evidence of her not getting wet. If Tamsin wants to avoid getting wet, and decides on a state that is a cause of not getting wet, she'll have made a decision that is *evidence* of an outcome she seeks. So she'll have made a good decision. If causal relations correspond to evidential relations in this way, knowing causal structure helps us make decisions that are evidence of outcomes we seek. So no wonder we care about causal relations.

Here's another example. Say Suzy is properly deliberating about whether to throw her rock. Her throwing her rock counts as a cause of the bottle's breaking if her deciding to throw (for the sake of the bottle's breaking) is good evidence of the bottle's breaking. Or consider scientists who investigate why certain stereoisomers (molecules that are right- or left-handed)

[5] I won't discuss how we individuate states of affairs, or how we distinguish causes from causal conditions. One could take the causal relata to be relativised to types or contrast sets, without significantly affecting the structure of the account. (Given how we deliberate and reason evidentially, my own preference is for relativising to types, but I won't defend that choice here.) The account could be extended to deliver probabilistic causation, but I'll only briefly consider how (n. 20).

are more prevalent in biological systems than their mirror equivalents by bombarding organic molecules with spin right- or left-handed electrons (Dreiling and Gay 2014). They find that left-handed bromocamphor is more likely to react with right-handed electrons (than left-handed), measured by the flow of bromide ions produced. By deciding whether to bombard left-handed bromocamphor with right- or left-handed electrons, a scientist can alter the ion current. The deliberative account correctly picks out the handedness of the electrons as a cause of the current intensity. A scientist's decision to use right-handed electrons on left-handed bromocamphor is good evidence for a higher current, when she deliberates on which type of electrons to use. So the electron-handedness is a *cause* of current intensity.

With these examples in mind, I'll now consider how the biconditional is to be read. Firstly, the biconditional involves a counterfactual on the right hand side. It makes a claim about what *would* be the case, *were* an agent to be properly deliberating. This is an important feature for securing the objectivity of causation (section 3.2). I discuss how such counterfactuals are evaluated in section 3.3.

What about evidential relations? A state of affairs is 'good evidence' for another if knowledge of its obtaining would, in general, license agents to infer to the latter and hold its obtaining fixed.[6] While what we are licensed to infer in any given case is sensitive our beliefs, we also have general evidential norms that pick out when knowledge of a state is typically license to infer to another (absent defeaters). It is these general evidential relations that the

[6] Note that good evidence need not license absolute certainty, or a degree of belief 1. It licenses certainty that rules out incompatible possibilities as 'serious possibilities' for the agent. For more on serious possibility, see Levi (1986, ch. 4).

deliberative account appeals to. For example, storm clouds looming may be good evidence of later rain, because knowledge of their presence would typically license an agent to believe there'll be rain (even if being told that these are merely thunder-clouds presaging a lightning storm would undercut this license). According to the biconditional, Tamsin's taking her umbrella is a cause of her staying dry if and only if her deciding to take her umbrella generally licenses belief that she'll stay dry.

Because the evidential biconditional appeals to general evidential relations, an agent's deciding is not simply evidence for that agent. If Tamsin's deciding is good evidence she'll take her umbrella, and you know of her deciding, you also have good evidence she will. Whether an agent *has* evidence depends partly on what observations she has made. But what her evidence is *for* does not depend on her background beliefs. Even if Tamsin has mistaken beliefs about the weather, storm clouds are still evidence of rain, and, if she's seen the storm clouds, she has evidence of rain. Good evidence also licenses agents to infer to the past *and* the future—a crucial feature if the temporal asymmetry of causation is to be explained, rather than presupposed. In section 3.2, I consider what relations might play the role of 'good evidence'.

The biconditional requires an agent's deliberation to be 'proper'. If an agent's deliberation is not proper, no causal relation is implied. Deliberation must satisfy certain epistemic requirements in order to be proper. These requirements reflect the fact that deliberation serves an epistemic function: it allows us to settle what we will do so we can engage in further planning (Bratman 1984). Serving this function requires us to be uncertain as we

deliberate.[7] While an agent might deliberate even if her evidence *does* settle what she will decide, she would be mistaken to do so. Her deliberation would not serve its normal function, and would not be 'proper'. Proper deliberation therefore requires: a) an agent's deciding on an option is good evidence the option obtains; b) her evidence leaves open which of several incompatible options she decides on; c) her evidence leaves open whether the state she decides for the sake of obtains. If Tamsin decides on taking her umbrella (for the sake of not getting wet) in proper deliberation, her deciding must be good evidence of her taking her umbrella and she must not otherwise have evidence that settles her taking it or her staying dry. If Tamsin's evidence already settles the day's being sunny, the biconditional does not imply that her taking her umbrella is a cause of her staying dry.

The requirements on proper deliberation derive from the epistemic function of deliberation. This fact distinguishes the deliberative approach from Woodward's. Woodward provides conditions on intervention without specifying how they matter to agents (section 1). The deliberative account, by contrast, bases its conditions on causation in a normative characterisation of deliberation. Because it matters to agents whether deliberation serves its epistemic function, it matters to agents whether their deliberations meet the conditions on proper deliberation.

The epistemic features of proper deliberation also explain an agent's apparent ability to intervene in the world without appealing to causal relations or beliefs. Agents appear free to decide in different ways because they are ignorant of *what* they'll decide and yet their decisions are evidentially relevant to how the world goes. The deliberative approach does not

[7] Related ignorance conditions are defended by Levi (1986, ch. 4) and Kapitan (1986).

merely presume agents are or appear free to intervene, or rely on causal characterisations of deliberation. This is an advantage it has over interventionist and reductive accounts, which at best merely presume agents are free to intervene (section 1).

The requirements on proper deliberation are what allow the biconditional to pick out causal relations. Evidential relations alone would get the wrong results. For example, say Tamsin has a strong habit of deciding on taking her umbrella whenever she sees rain. If so, her deciding on taking her umbrella is good evidence of rain. Yet her taking her umbrella isn't a cause of rain. The deliberative approach gets the right result here. If Tamsin's seeing the rain already settles there being rain, her deliberation is not proper. So a causal relation is not implied. In section 4, I'll offer a general explanation of why the evidential biconditional does not count spurious correlations as causal.

2.2 Correspondence without Reduction

The evidential biconditional claims a correspondence between evidential and causal relations. But there are different ways this biconditional might be read. It might seem that the biconditional tells us about causal reasoning, or the *concept* CAUSATION. Perhaps it analyses the concept, or tells us how agents employ the concept. This is *not* how the biconditional is to be read. The biconditional makes claims about causal relations themselves. It provides necessary and sufficient conditions on causal relations, distinguishing them from non-causal relations. Because the biconditional is about causal relations, it can explain why we should care about them, in contrast to the all the other relations we might pick out as causal. When we decide on causes for the sake of their effects, we decide on states of affairs that are good evidence for outcomes we seek.

Because the biconditional makes claims about causal relations, it provides a constraint on first-order metaphysical accounts of causation. Whatever relations an account picks out as causal must satisfy the evidential biconditional.[8] The biconditional is a condition of adequacy of other accounts. If an account satisfies the biconditional, it can use the biconditional to explain the usefulness of causation. The deliberative approach can therefore justify other metaphysical accounts, and show why the relations picked out deserve to be called 'causal'.

These metaphysical upshots of the deliberative approach distinguish it from other agent-based accounts. Huw Price argues for a related biconditional on causation: 'An event *A* is a cause of a distinct event *B* if and only if ensuring that *A* rather than not-*A* would be an effective means–end strategy for a free agent whose overriding desire is that it should be the case that *B*' (1991, p. 170)—see also Price (2012, p. 509). While this might suggest a metaphysical account of causation, perhaps a response-dependent account in which 'causal relations are…mind (or more particularly agent) dependent, just as secondary qualities are sensory agent dependent' (1991, p. 173), more often Price brackets the nature of the account (2012, p. 488) or takes himself to be concerned with the concept CAUSATION—perhaps providing an 'analysis' (1991, p. 159), conditions for understanding the concept (1993, p. 201), an account of its function, or a genealogy (1992, p. 514). But an account that is merely about the concept CAUSATION, doesn't address Russell's challenge. It doesn't explain how causation itself fits into the picture of the world presented by fundamental physics.

The deliberative account does not aim to reduce causal relations to evidential relations. But it

[8] The deliberative approach can tolerate some exceptions to the biconditional. Provided the biconditional is satisfied in sufficient paradigmatic cases, it can still explain why causation matters to us in general.

does aim to derive causal structure from evidential structure. It ultimately aims to relate causation to fundamental laws using evidence as a half-way step (see figure 1).

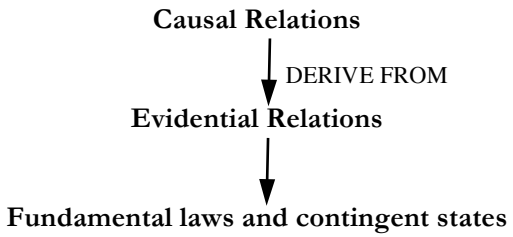

Figure 1: How the deliberative approach relates causation to fundamental laws.

For example, say Suzy's deciding to throw her rock is good evidence of the bottle breaking. Presumably this is because in the circumstances in which Suzy throws the rock (in a given direction), the underlying dynamical laws and an initial probability distribution make it highly probable that the bottle breaks. (More details on this in section 3.2.) If evidence relates to fundamental laws in this way, the approach provides an even stronger answer to Russell's challenge. But the approach does not rely on evidence being more *metaphysically* fundamental than causation: it may simply be that we can more easily relate causal relations to fundamental laws and probabilities via evidential relations.

The deliberative approach does its explanatory work independently of settling the metaphysical nature of the causation. This is an aim it shares with interventionist accounts. The deliberative approach uses a correspondence between evidence and causation to explain why we should care about causal relations. A more radical stance would be to argue that causation has no metaphysical nature. I won't argue for this stance. It is enough for my purposes that the biconditional is what explains why we should care about causal relations

and why causes come before their effects (section 4).

Altogether, the deliberative approach sits somewhere between perspectival accounts like Price's, interventionist accounts, and reductive accounts. In common with reductive accounts, the approach relates causation to fundamental laws. But unlike reductive accounts, it does not merely aim to give necessary and sufficient conditions on causation. It explains why causation matters. In common with interventionist account, the deliberative approach takes causation to be centrally about manipulation and control. But unlike interventionist accounts, it uses an evidential characterisation of deliberation that does not rely on causal notions, and appeals to the epistemic function of deliberation to explain why satisfying the relevant conditions matters to agents. In common with Price's account, the approach thinks about causation via its relevance for deliberation. But, unlike Price's, it gives a substantive condition on causal relations themselves and fits them into a physical picture of the world.

3. Beyond Subjectivism

The deliberative approach ties causation to deliberation. For this reason, it might seem to imply a form of subjectivism, and be limited to situations involving agents. In this section, I explain how the deliberative approach delivers objective causal relations, and how it differs from other agent-based accounts.

3.1 Aren't Agents Causal?

Appealing to deliberation to give an account of causation might seem like the wrong approach to take. After all, deliberation and agency are causal phenomena. How can we get a substantive constraint on causation if causal notions are already involved? Similar worries

have been raised against agent-based accounts that take **A** to cause **B** if and only if bringing about **A** is a means of bringing about **B**. It seems they will be circular and uninformative because 'bringing about' is a causal notion (Hausman 1997; Woodward 2003, pp. 123ff).

Price defends an agent-based account that is particularly vulnerable to this concern. He claims that agents must take their acts to be *caused* by their deliberation alone as they deliberate. 'To introduce the agent is in effect to assume an independent causal history to the event *A*. Those probabilistic correlations that survive this assumption seem to have claim to be counted as genuine effects of *A*' (1991, p. 169)—see also Price (ibid. pp. 165–6, 1986, pp. 199–201, 1993, p. 261).[9] Even though a circular account can be illuminating, an account will be more informative if it does not appeal to causal notions. The deliberative approach characterises deliberation in evidential and non-causal terms. Neither we as theorists nor the deliberating agent need to make explicitly causal assumptions.

One might also worry that evidential relations are themselves causal. In the next section, I outline an account of evidential relations that doesn't appeal to causation. But even if we do characterise evidence in causal terms, we still have a sufficiently independent grasp of the concept EVIDENCE for the biconditional to be explanatory and illuminating. Evidence plays a distinct normative role in empirical enquiry. Whatever evidential relations turn out to be, they have been picked out as evidential relations *because* they are good ways for us to reason from one state of affairs to another. Appealing to evidence provides a non-trivial explanation of why causal relations are useful.

[9] Price later (2012, pp. 528–31) follows Ismael (2007) in tracing this assumption back to a special epistemic authority agents have over their own decisions—no matter what an agent's evidence about what he will decide, he can always trump this evidence and decide otherwise. But this apparent *ability* to trump evidence remains an unexplained primitive. For further details, see Fernandes (2016).

It's also worth keeping in mind the particular roles the biconditional plays in the deliberative approach. The biconditional gives necessary and sufficient conditions on causation, relates causation to evidence, and constrains metaphysical accounts. Playing these roles does not require evidence to be non-causal or more metaphysically fundamental than causation. The biconditional also explains why we should care about causation. Provided we have a sufficiently independent grasp of evidence, as I have argued we do, this explanation can be illuminating, even if we also give a causal account of evidence. The explanation of causal asymmetry (section 4), moreover, does not rely on causal features of evidence.

3.2 Objective Causal Relations

Does the deliberative approach imply that causation is subjective, and depends on agents' beliefs and abilities? Woodward raises this as an objection against agent-based accounts. He claims they won't deliver causal relations that are suitably objective (2003, 118ff.). One way causation could turn out to be subjective is if causal relations only obtain *when* agents are deliberating. But the evidential biconditional is explicitly counterfactual: **A** is a cause of **B** if and only if *were* an agent to decide on **A** for the sake of **B** (in proper deliberation), her deciding on **A** *would be* good evidence of **B**. (I consider how to evaluate these counterfactuals in section 3.3.) Agents need not actually deliberate or even be able to properly deliberate on **A** for a causal relation to hold. So causation does not depend on our practical or technological abilities to bring about **A**. Another way causation might be subjective is if it depends on an agent's beliefs. But the deliberative approach uses *objective* evidential relations, rather than subjective beliefs. Independently of what an agent believes or desires about her decisions, her decisions are evidence for some states and not others. Even if Tamsin believes

or desires that her deciding on sunshine settles there being sunshine, this does not make it so. While we need to consider deliberating agents in order to understand the relevance of causation, causation does not depend on agents' beliefs.

Other agent-based accounts rely more heavily on the deliberators' beliefs. Price argues that causation is 'perspectival', and that causal relations correspond to the evidential relations that hold 'from the free agent's distinctive point of view' (1991, p. 173) or when 'assessed from the agent's distinctive epistemic perspective' (2012, p. 494). According to Price, it is only when we take up the point of view of a deliberating agent, with its distinct pattern of subjective probabilities, that causation appears in our world-view. But this perspectivalist approach does not directly answer Russell's challenge. It doesn't explain how causation itself fits into a world-view, but rather gives subjective conditions on when we use causal concepts. The deliberative approach is not perspectivalist in Price's sense. Causal and evidential relations appear from outside the deliberator's point of view. Say Tamsin's decision to take her umbrella is good evidence of her not getting wet. If so, it is not just evidence for her. Tod, learning of her decision, also has evidence that Tamsin won't get wet. He has this evidence in virtue of observing her behaviour, and reasoning theoretically about it, not in virtue of taking up her perspective. The conditions on proper deliberation guarantee her decision will be evidentially relevant for Tamsin. But her decision still counts as evidence for Tod.

What are the objective evidential relations the deliberative approach appeals to? One possibility is that they are evidential probabilities that derive from shared norms of empirical enquiry, similar to Gillies' intersubjective probabilities (2000). Another is that they are causal

relations. These alternatives could be used to justify a more objective version of Price's account or non-reductive interventionist accounts, respectively. A more interesting alternative, however, is to take evidential relations to be physical chances. This option secures a stronger connection between fundamental laws and causation, more fully answering Russell's challenge. Existing accounts of chance that relate chance to fundamental laws include Lewis' 'best systems' account (1986) and statistical-mechanical accounts (Albert 2000, ch. 6, 2015, ch. 1; Loewer 2007). Lewis' account, however, builds in a temporal asymmetry by hand. I end this section by outlining an account of chance based on statistical mechanical approaches.

Say Suzy picks up her stone and decides on throwing for the sake of the bottle breaking. If the fundamental laws are deterministic, a full specification of the state of the world will fix whether the bottle breaks or not. But agents like Suzy won't typically have access to the full state of the world. These relations aren't the evidential relations we use in higher-level science and when we reason. A more plausible starting point is to think that agents have access to the macrostate of the subsystem of Suzy, the stone and the bottle, and assume the subsystem is embedded in a stable and familiar surrounding environment. Given such assumptions, can fundamental laws determine the chance of the bottle's breaking?

A promising approach is to make use of the explanatory resources within science to work out these chances. Fundamental physics aims to explain the success of higher-level sciences, including non-fundamental physics. Given that chances feature in higher-level science, *scientific* explanations should be able to explain higher-level chances in terms of fundamental laws and additional features. We have toy examples where this is possible. Say there is a box

containing gas particles that are partially dispersed at time t_2. Given the fundamental laws, the standard statistical-mechanical probability distribution, the macrostate of the box, and the fact that the box is not further interfered with, we can derive the result that (with high probability) the gas will be more dispersed at a later time, t_3.[10] So the partially dispersed gas is evidence of future dispersion. The same procedure won't work going backwards in time—it will deliver the result that the gas was *more* dispersed in the past, at t_1: not the result we observe. But if we also conditionalise on the system beginning in a low-entropy state at t_0 (a standard move in Boltzmannian statistical mechanics), it's highly probable that the gas was less dispersed at t_1. So partially dispersed gas is evidence of earlier clustering.

Why think these kinds of explanations are available more generally? Firstly, we have arguments that for folk physics to get going, systems have to be sufficiently often in quasi-isolation from one another (Elga 2007). The scientists investigating the decay of bromocamphor have to be able to isolate the ion current from other things that could alter its flow, other than the incoming electrons. Secondly, for us to do experiments in fundamental physics, or for fundamental physics to explain the success of other sciences, fundamental laws have to explain and be derivable from regularities that appear at the macroscopic level—concerning instrument knobs, rocks, CO_2 emissions, and so forth. Thirdly, we often assume systems are in normal or default conditions—a fact appealed to in a range of accounts of causation, particularly counterfactual accounts (Hart and Honoré 1985; Menzies 2007; Hitchcock 2007; Paul and Hall 2013, pp. 49−53). We assume stable background conditions, and regular behaviour of systems and subsystems, to distinguish

[10] For details of this Boltzmannian approach and related proposals, see Horwich (1987, ch. 4), Albert (2000, ch. 3; 2015 ch. 1), Loewer (2007), and references therein. Note that the approach involves introducing a probability distribution at the fundamental level (even if the laws are deterministic), and so does not in itself provide a reduction of chance.

causes from causal conditions and to derive particular token-causal judgments.[11] In the case of Suzy, we assume there aren't competing rock throwers like Billy around. Assumptions like these, however they are justified, underlie our evidential reasoning in scientific and everyday life as we reason from one system's state to another's. They can be used to provide the evidential chances needed for the deliberative approach.

3.3 Counterfactuals

I've argued that evidential relations can deliver objective causal relations and relate them to fundamental laws. But there's a worry. The evidential biconditional refers to what an agent's decisions *would be* evidence for, *were* an agent to be properly deliberating. These counterfactuals are needed for the account to deliver all the causal relations there are. But can we evaluate such counterfactuals without presuming causal notions? For example, does the deliberative account imply that volcanoes erupting or earthquakes occurring are the causes of their effects, given agents never properly deliberate on these states?

There's a sense in which this worry is less pressing than under reductive accounts. The motivation for the deliberative approach is to make sense of why we care about causal relations. The approach is committed to cases where agents *can* deliberate on causes being central and paradigmatic. Our interest in causation in cases where we can't derives from our interest in cases where we can. While we can extend causal notions to situations we don't control, the less grip our notions of control have, the less grip our notions of causation will

[11] The deliberative approach faces the same concerns with pre-emption as other counterfactual accounts (Hall 2004; Paul and Hall 2013, ch. 3). My preferred response is to appeal to assumptions about normal conditions—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 we needn't aim to give complete conditions for particular token causal judgments. I agree with Hitchcock (2007, p. 511−2) and Woodward (2003, p. 85) that our intuitive causal judgments aren't as central to causation as the patterns of counterfactual dependence.

have. Ultimately the account may want to explain away some of our causal intuitions. But the account should at least be able to evaluate counterfactuals that aren't so 'distant' from ordinary cases.

Sometimes we can evaluate counterfactuals concerning a deliberating agent's evidence by adding an agent to a situation or adding or subtracting other features so that an agent can properly deliberate. We can do this when these changes don't alter the relevant dynamics of the relevant subsystem.[12] For example, say there are no agents near Suzy's bottle, but a flying rock causes it to break. If adding in Suzy doesn't generally alter how a subsystem of a rock and bottle behaves after the rock is launched, we can construct a situation in which Suzy is present and properly deliberates on throwing a rock. Whether her decision is evidence for the bottle breaking then determines whether the rock's motion (in the original system) caused the bottle's breaking. Or say a set of billiard balls is encased in unbreakable glass. If removing the glass case doesn't alter how the balls behave once they're set in motion, we can construct a situation in which the glass is removed and an agent properly deliberates on their movements, to determine what causal relations there are in the original system. This approach relies on the fact that changing features doesn't disrupt all evidential relations in quasi-isolated subsystems.

But there are cases where adding an agent would disrupt the dynamics of the system. For example, say a symmetry-breaking state in the early universe caused the current universe to be mostly matter rather than antimatter.[13] Constructing a similar situation with a deliberating agent present would require making significant changes to the dynamics of the system.

[12] Evidential relations to the past (or the future via the past) can be altered, for reasons given in section 4.
[13] My thanks to Michael Hicks for the example.

Significant changes may also be required for agents to deliberate on volcanoes erupting or earthquakes occurring. There are also states that it is logically impossible for an agent to properly deliberate on—such as whether any agents ever exist.[14]

In these cases, a hypothetical non-physical agent can be introduced to allow us to evaluate the relevant counterfactuals and work out what evidential relations would hold for a properly deliberating agent. By stipulation, the presence of a hypothetical agent has no physical effects on the system. Even though none of us are hypothetical agents, the evidential structures relevant for hypothetical agents are of the same form that actual agents make use of. They are structures where even though available evidence doesn't determine how the system goes (so an agent can properly deliberate), states that are evidentially available to decide on are evidence of further states of interest. This strategy does justice to the initial motivation for the approach—causal relations are useful because they're of an evidential form that agents can generally make use of in deliberation.

For example, say the early state of the universe doesn't determine whether a particular symmetry-breaking state (**E**) obtains, or whether a further state (**F**) obtains—say, the universe being mostly matter. Stipulate that a hypothetical agent's decision on **E** would evidentially settle **E** (and likewise for not-**E**). A hypothetical agent can then properly deliberate on **E** for the sake of **F**. If **E** is good evidence for **F** (something we might determine independently), then the hypothetical agent's decision on **E** for the sake of **F** (in proper deliberation) would be good evidence of **F**. So **E** counts as a cause of **F**.[15] We use our

[14] My thanks to a referee at this journal for the case.

[15] This approach might seem to deliver far too many relations as causal. My general response to backwards and spurious causation comes in section 4. Once this response is in place, backwards causation is ruled out in these

knowledge of the evidential relation between **E** and **F** to work out what causal relations obtain. Even though no agent can properly deliberate on **E**, the hypothetical agent is making use of the same general form of evidential structure that would matter to deliberating agents (were they able to deliberate).

Here's another example. Say the state of the early universe does not determine whether any physical agents ever exist (**A**), or whether tools exist (**T**). We can introduce a hypothetical agent who properly deliberates on **A** for the sake of **T**. Say **A** is evidence of **T**. Then the hypothetical agent's decision on **A** for the sake of **T** would be evidence of **T**, and **A** counts as a cause of **T**—even though, for logical reasons, no actual physical agent could ever properly deliberate on the existence of physical agents.

There are more difficult cases where there is no pre-existing evidential gap in the system. A gap must then be opened up by breaking evidential relations between current states and previous states of the system, while not disrupting evidential relations between present states and future states (except by way of the past). In the case of the early universe, if the state of the early universe *does* settle **E** obtaining, not-**E** is made evidentially available by breaking evidential relations between **E** and previous states, and not otherwise interfering with evidential relations between **E** (or not-**E**) and later states. Causal relations are then determined as earlier. The temporal asymmetry introduced reflects a temporal asymmetry of agency explained in section 4.2. This procedure may look somewhat artificial. But this

---

counterfactual cases by projecting aspects of our asymmetric structure as agents onto the hypothetical deliberator. Does this appeal to projection introduce an undesirable subjectivism into the account? Not if the asymmetries of agents are explained in objective asymmetric terms (section 4.2). In simple worlds lacking these asymmetries, there may be no reason to project in one way rather than another. If we still feel that causation is temporally asymmetric at such worlds, this feeling should be explained away as a mere *habit* of projection.

shouldn't be surprising. As counterfactual cases become less like those that we deliberate in, the method for determining causal relations becomes less natural to apply.

4. Causal Asymmetry

A notable feature of causation at our world is that causes always come prior in time to their effects. While we might accept backwards causation at the microscopic level or in other metaphysically possible worlds, we don't ordinarily encounter macroscopic causes coming before their effects. Yet the laws of fundamental physics don't distinguish between the past and future (or at least not so as to readily explain causal asymmetry). Causation's temporal asymmetry needs to be explained if we're to fit causation into a scientific picture of the world. A final advantage of the deliberative approach is that it can do just that.

4.1 Correlations to the Past

Initially it might seem that the deliberative approach is ill-suited to explaining 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 an evidence-based approach will imply, incorrectly, that causation is directed towards both the past and the future.[16]

To determine whether the evidential biconditional implies backwards causation, we need to consider what evidential relations there actually are between a properly deliberating agent's decisions and past states. Let's return to Tamsin and her umbrella. Say Tamsin usually takes her umbrella when she sees rain, so that her deciding to take her umbrella is, in ordinary

[16] There is also a concern with counting spurious correlations (such as joint effects of a common cause) as causal. My response applies equally to both concerns.

cases, good evidence of rain. It may seem that the evidential biconditional will imply that Tamsin's taking her umbrella is a *cause* of rain. But say Tamsin hasn't yet seen the weather. Would her decision to take her umbrella count as evidence of rain? No. The reliability of the evidential connection between her decision and the rain depends on her seeing the rain, or otherwise having evidence of it. Without her having this evidence, we have no reason to think there'll be a correlation between her decision and the rain—Tamsin would be merely lucky if she decided to take her umbrella just in cases where there was rain.

What this implies is that even if Tamsin's deciding to take her umbrella is ordinarily evidence of rain, this is not the case if she decides to take her umbrella *for the sake of there being rain* in proper deliberation. Why? Because proper deliberation requires Tamsin not to have evidence that settles the state she's deciding for the sake of. So she can't have evidence of rain. But without having evidence of rain, her decision is not good evidence of rain—and so her taking her umbrella doesn't count as a cause of rain. Note that it's not only from Tamsin's perspective that there is no evidential correlation between her decision and the rain. If Tamsin's housemate Tod uses Tamsin's behaviour to decide whether *he* should take his umbrella, he would do well to ignore Tamsin's behaviour when she properly deliberates—her decision would not be evidence of rain, even for him.

In the case where Tamsin's decision is good evidence of rain, the evidential correlation goes *via* her observing the weather. Might this case generalise? Evidential decision theorists have attempted to show that evidential relations between past states and acts are *always* mediated by other states that form part of deliberation, such as beliefs, desires and judgements. If an agent believes they are in such a deliberative state, and that this deliberative state is

correlated with the past, they already have (subjective) evidence of the past. This 'screens off' the (subjective) evidential relevance of their act for the past state—what act they decide on is no longer evidence for the past. So if agents decide on acts that are evidence of good outcomes, they will not be led astray and act for the sake of the past.[17]

For example, say Gina knows that a gene is the common cause of both smoking and cancer, (figure 2) and deliberates about whether to avoid smoking to avoid having the gene.

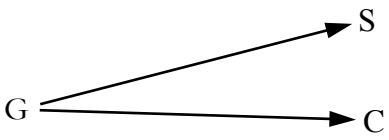

Figure 2: A gene (**G**) as a common cause of smoking (**S**) and cancer (**C**).

If the correlation between having the gene and smoking is mediated by a desire to smoke, Gina's belief that she has this desire and that it is correlated with the gene gives her (subjective) evidence that she has the gene. This evidence 'screens off' the (subjective) evidential relevance of smoking for having the gene. No matter what she decides at this point, her smoking or not gives her no additional information about whether she has the gene. So she may as well smoke if she wants to. If agents always have the relevant beliefs about such deliberative states, the correlation between their decision and the past state will always be screened off.

[17] This kind of screening off is considered in Nozick (1969), and appealed to by Eells (1982, 1984), Jeffrey (1981), Horgan (1985), Price (1986, 1991, 1993) and Horwich (1987, ch. 11). I consider concerns below. Additional concerns are raised by Sobel (1994) and Papineau (2001).

If this defence of evidential decision theory is successful, then the evidential biconditional will not imply backwards causation. An agent's deliberation is proper only if her objective evidence leaves open whether the state she's deciding for the sake of obtains. If (objective) evidential relations towards the past are always mediated by beliefs, desires, or other states that form part of an agent's evidence while she is deliberating, these provide evidence of the past that make deliberation on the past improper (figure 3). When an agent's deliberation is proper, this must be because the mediating state is *not* part of her evidence. But then there is no reason to suspect an (objective) evidential correlation between her deciding and the past state. This is the relation needed for causation. So such cases do not imply backwards causation.[18] Note that this explanation does not appeal to causal notions, in contrast with Price (1986, p. 201; 1991, p. 166; 1993, p. 261).

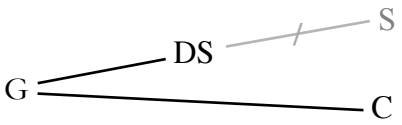

Figure 3: The evidence Gina has while deliberating on smoking (**S**) for the sake of not having the gene (**G**), given the causal structure in figure 2. Gina has evidence of the deliberative state (**DS**) that already settles whether she has the gene, and so whether she will get cancer (**C**), making her deliberation improper (even if it doesn't settle her smoking).

4.2 Objections and Replies

A number of objections have been raised against the screening off defence of evidential decision theory. I'll briefly consider five. My strategy will be to show that what's needed to

[18] Isn't Tamsin's deciding evidence she'll take the *means* to her end? Her deciding to get her umbrella might be good evidence she'll walk down the hall to it. Won't the deliberative account imply her getting her umbrella is a cause of her walking? Only if Tamsin decides to take her umbrella *for the sake of* walking through the hall, and deciding in such a way is instrumentally irrational, a sign of motivational deficiencies (perhaps Tamsin needs an excuse to exercise). It is not how we should deliberate.

defend the deliberative approach to causation is *less* than what's needed to defend evidential decision theory. As far as evidential decision theory remains defensible, so does the deliberative approach.

Objection 1: 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 a causal relation is not implied, because the decision does not settle the past state.[19]

Objection 2: What if the correlation does not go via a deliberative state? Perhaps Gina has an addictive tendency, and smokes regardless of her beliefs and desires.

For a relation between **A** and **B** to be causal under the deliberative approach, it must be the agent's *deciding* on **A** that is correlated with **B**, not **A** itself.

Objection 3: What if the deliberative state isn't present at the start of deliberation, but obtains later?

The deliberative account does not require there to be a particular point in deliberation by which screening off has occurred and a decision is recommended. What is required is that when the agent *decides*, the deliberation was proper and the decision has the right evidential relations.[20]

---

[19] If one generalises the account to deal with probabilistic causation, a different response is needed. Say **A** is causally *relevant* to **B** if and only if deciding on **A** *changes the probability* of **B**, relative to not deciding on **A**. In this case, the mediating state provides as much evidence for the past as the decision does—so the evidential relevance of the decision for the past is still undermined.

[20] Jeffrey (1981, 1983) gives a 'ratificationist' defence of evidential decision theory along these lines. But this defence faces problems of its own—see Horwich (1987, pp. 188−9).

Objection 4: What if the correlation between the decision and past state holds, regardless of whether any particular mediating state occurs during deliberation? If so, there is no screening off. Deliberation can remain proper, even while the decision is evidence for the past.

This is effectively what happens in traditional Newcomb cases, where an infallible predictor arranges the past to match whatever decision he predicts an agent will make (Nozick, 1969). But ruling out backwards causation in these cases is not required to defend the deliberative approach. The defence only needs to show that causation does not run backwards in realistic cases we encounter (so-called 'medical' Newcomb cases), not unrealistic cases in which the epistemic capabilities of the predictor go far beyond those used in ordinary evidential reasoning.[21]

Objection 5: The screening off defence requires agents to have too much self-knowledge. Actual agents may not realise they're in the relevant deliberative state or not recognise it's relevance for the past. So there will be no screening off (Skyrms 1980, p. 131; Lewis 1981, pp. 10−11).

The deliberative approach appeals to *objective* evidential relations, and what evidence an agent *has*, rather than the *subjective* probabilities and beliefs of evidential decision theory. Gina's deliberative state may give her objective evidence of the past, even if she does not believe it does. Nor does she need to believe that she's in a deliberative state for it to provide her evidence. Provided these states are accessible to her, they plausibly count as part of her evidence. Agents may fail to live up to these objective evidential standards. But these standards are reasonable given that we do aspire to self-reflectively respond to objective reasons and evidence in deliberation.

[21] More strongly, it's plausible that traditional Newcomb cases *do* involve backwards causation (Price 2012; Nozick 1969, p. 134). The abnormal evidential structure implies abnormal causal structure.

There are further concerns to be faced in defending the deliberative approach. If agents are never able to decide based on their beliefs, then a causal relation that corresponds to objective evidential relations will not be so useful. But 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 too demanding. Overly high standards would risk making all deliberation improper (in addition to ruling out backwards causation). Let me point to two considerations that suggest the standards needed don't rule out deliberation in general. Firstly, we don't think agents are systematically insensitive to their evidence when they deliberate. If so, their actual deliberations on the future suggest they satisfy conditions of proper deliberation. Secondly, agents are complex and self-reflective. Their beliefs and desires can be further inputs to deliberation. For this reason, a deliberative state that provides evidence of the past is often not enough to settle what an agent decides. So agents can properly deliberate on the future, even while backwards causation is ruled out.

4.3 Deciding on the Future

To summarise so far, here is how the deliberative approach explains causal asymmetry. If decisions are reliably correlated with past states, this goes via agents being in particular deliberative states at times in between, states that give them evidence of the past. Their having this evidence makes their deliberation on the past improper. It is only when agents aren't in such states that deliberation is proper. But then agents' decisions aren't evidence for the past. The situation towards the future is very different. Correlations between decisions and future states are not mediated by deliberative states. So when deliberation on the future is proper, there can still be evidential relations between an agent's decisions and states of

affairs further in the future. So there can be forwards causation. Putting these two pieces together, causal asymmetry is due to the fact that deliberative states mediate correlations towards the past but not the future. If we assume that states of the agent mediate evidential relations between their decisions and states at other times, it is the fact that deliberation comes *before* decision that explains causal asymmetry. A final step in the explanation is to explain why this asymmetry holds: why deliberation comes before decision. This is a significant improvement on assuming the deliberative asymmetry as a primitive (Price 1993, p. 260).

To explain why agents deliberate before deciding, rather than after, we should appeal to the epistemic characterisation of deliberation (section 2) and an epistemic asymmetry. Begin with the fact that we have memories of the past but nothing like memories of the future. In ordinary cases of deliberating, agents know about the immediate past in their vicinity, independently of what decisions they make now. As Tamsin deliberates on whether to take her umbrella, she is sure she was walking through the hall a minute ago. 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 there as she does. But we have knowledge of the future by way of our current decisions, intentions and habits (in the first case), or more general beliefs about the present or past and reliable behaviour (in the second). Tamsin's certainty of her own immediate past is not dependent in this way—independently of her present decisions, intentions, habits or beliefs about her present surroundings, she has *memories* of her past. She does not have anything like *memories* of her future.

According to the epistemic characterisation of deliberation, an agent can't deliberate if she is

certain how she'll decide. 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. Her future remains suitably uncertain. As far as an agent's deliberative structure is generic, and tracks general epistemic features, there will be a temporal asymmetry of deliberation.[22] It is because we have memories of the past, but nothing like memories of the future, that we deliberate towards the future and not the past. Note that this explanation is not available on competing models of deliberation (Ismael 2007; Price 2012) that deny ignorance conditions on deliberation.

Can we explain this epistemic asymmetry in turn? Albert (2000, ch. 6; 2015, ch. 2) gives an explanation of an equivalent asymmetry of records. Records, like memories, are local states of the present that provide reliable evidence of states of systems at other times, beyond generic regularities, and independently of knowing what happens to the system between now and then. Albert argues that the same macroscopic constraint on the early state of the universe that explains the asymmetry of entropy, the Past Hypothesis, (and the lack of such a constraint on the future) explains why there can be records of the past (and not the future). Even though Albert uses this asymmetry of records to defend a statistical-mechanical account of causation, his explanation of it proceeds in non-causal terms, and so can be used to defend the deliberative approach. So the asymmetry of agency can be explained in objective physical terms.[23] The deliberative approaches traces causal asymmetry back to the

[22] What if there are past decisions you don't remember? Can you then deliberate? Ignorance conditions won't rule out such deliberation. But if your deliberation now is to have an evidential bearing on your past decision, your past decision must settle how your deliberation will go, independently of what happens to you at times in between (when you lack evidence). Your decision will function as a record of the future. And we don't think there are such things.

[23] Moreover, even if Albert's account can't be used to explain the asymmetry of agency, we can expect some other ultimately physical account to do so.

same temporal asymmetries that give rise to asymmetries of agents (figure 4). But agency still plays an essential role in explaining why correlations towards the past are undermined, and so why causes always comes before their effects.

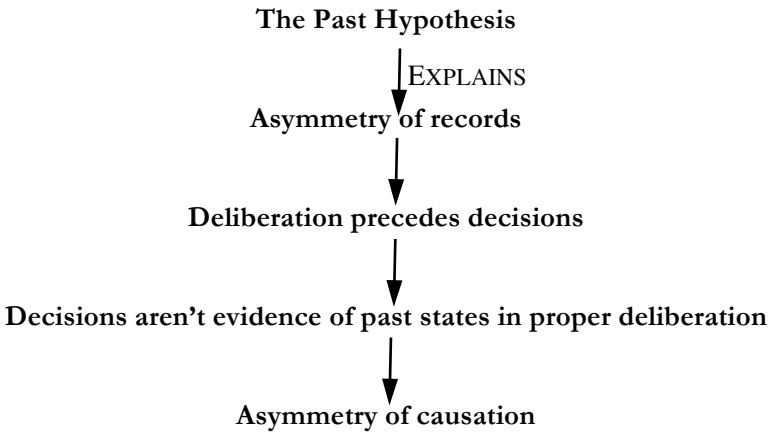

Figure 4: The explanation of causal asymmetry under the deliberative approach.

5. Conclusion

Causation might have seemed like a deep fundamental feature of the world, something whose importance we could make sense of independently of ourselves. But we should make sense of it by reference to us. According to the deliberative approach, causal relations correspond to the evidential relations we use in deliberation. Causal relations obtain just when decisions for the sake of outcomes we seek (in proper deliberation) would be good evidence of those outcomes. This correspondence explains why we pick out particular relations as causal: they are needed for good decision-making. Even though causal relations are crucially relevant for deliberation, however, they do not depend on our perspectives, beliefs, desires or practical abilities. Causal relations are as objective as the evidential relations that underlie good reasoning. This approach strikes the right balance between capturing the relevance of causation for deliberation, and its objectivity.

The deliberative approach also explains the temporal asymmetry of causation—an asymmetry not reflected in the fundamental laws. When agents properly deliberate about what to do, their decisions aren't evidence for any past outcomes they may seek. From this it follows that causes always come before their effects. This asymmetry in what an agent's decisions are evidence for can ultimately be explained in physical terms. Once again, even though causal asymmetry should be explained by reference to deliberation, causal asymmetry is not merely perspectival or up to us.

This approach demonstrates a broad strategy for doing scientifically informed philosophy: use agential standards to pick out objective relations, explain their temporal features, and reconcile them with fundamental physics. In this particular case, the deliberative approach relates evidential and causal relations, via deliberation. The account identifies, at the most general level, how the evidential structure of the world relates to its causal structure. By doing so, it fits causation into a scientific picture of the world.[24]

References

Albert, David. Z. 2000. *Time and Chance*. Cambridge, Mass.: Harvard University Press.
———. 2015. *After Physics*. Cambridge, Mass.: Harvard University Press.

[24] My heartfelt thanks to Achille Varzi, David Albert and Jenann Ismael for their inspiring conversations, incisive criticisms and unwavering support throughout this project, as well as to Barry Loewer, Wolfgang Mann, Huw Price, John Collins, Matthew Simpson and Arif Ahmed for their careful guidance, ideas and critiques. I'd also like to thank the following for even more in the way of helpful suggestions and encouragement: Borhane Blili Hamelin, Michael Hicks, Shyane Siriwardena, Jorge Morales, Zac Al-Witri, Jonathan Fine, Georgie Statham, Brad Weslake, Mathias Frisch, Sidney Felder, Thomas Blanchard, Heather Demarest, John Maier, Neal Groothuis, Christian Loew, David Braddon-Mitchell, Kristie Miller and Hugh Mellor, as well as audiences at Columbia University, the University of Cambridge, the University of Sydney and the University of Arizona.

Dreiling J. M. and Gay, T. J. 2014. Chirally Sensitive Electron-Induced Molecular Breakup and the Vester-Ulbricht Hypothesis. *Physical Review Letters* 113(11): 118103.

Eagle, Anthony. 2007. Pragmatic Causation. In Price and Corry, 156−190.

Earman, John. 1976. Causation: A Matter of Life and Death. *The Journal of Philosophy* 73(1): 5−25.

———. 1986. *A Primer on Determinism*. Dordrecht: Reidel.

Eells, Ellery. 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.

———. 2007. Isolation and Folk Physics. In Price and Corry, 106–119.

Fernandes, Alison. 2016. Varieties of Epistemic Freedom. *Australasian Journal of Philosophy* 94(4): 736–751.

Field, Hartry. 2003. Causation in a Physical World. In Michael J. Loux and Dean W. Zimmerman (eds.) *The Oxford Handbook of Metaphysics*. Oxford: Oxford University Press, 435–60.

Frisch, Mathias. 2010. Does a Low-Entropy Constraint Prevent Us from Influencing the Past? In Gerhard Ernst and Andreas Hüttemann (eds.) *Time, Chance and Reduction*. Cambridge: Cambridge University Press.

Gillies, Donald. 2000. *Philosophical Theories of Probability*. London: Routledge.

Hart, H. L. A. and Honoré, A. 1985. *Causation in the Law*, 2nd edn. Oxford: Oxford University Press.

Hausman, Dan. 1997. Causation, Agency, and Independence. *Philosophy of Science* 64(4 Supplement): S15−S25.

Hall, Ned. 2004. Two Concepts of Causation. In John Collins, Ned Hall and Laurie Paul (eds.) *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press, 225–76.

Hitchcock, Christopher. 2007. Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review* 116(4): 495−532.

Horgan, Terry. 1985. Counterfactuals and Newcomb's Problem. In Richmond Campbell and Lanning Sowden (eds.) *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*. Vancouver: University of British Columbia Press, 159−182.

Horwich, Paul. 1987. *Asymmetries in Time*. Cambridge Mass.: MIT Press.

Ismael, Jenann. 2007. Freedom, Compulsion and Causation. *Psyche* 13(1): 1−11.

Jeffrey, Richard. C. 1981. The Logic of Decision Defended. *Synthese* 48: 473−492.

Kapitan, Tomis. 1986. Deliberation and the Presumption of Open Alternatives. *The Philosophical Quarterly* 36(143): 230−251.

Levi, Isaac. 1980. *The Enterprise of Knowledge*. Cambridge Mass.: MIT Press.

———. 1986. *Hard Choices*. Cambridge: Cambridge University Press.

———. 1990. Chance. *Philosophical Topics* 18: 117–149.

Lewis, David. 1973. Causation. *Journal of Philosophy* 70: 556–67.

———. 1979. Counterfactual Dependence and Time's Arrow *Noûs* 13: 455–76.

———. 1981. Causal Decision Theory. *Australasian Journal of Philosophy* 59: 5–30.

———. 1986. A Subjectivist Guide to Objective Chance. In *Philosophical Papers: Volume II*. New York: Oxford University Press, 114–132.

Loewer, Barry. 2007. Counterfactuals and the Second Law. In Price and Corry, 293–326.

Menzies, Peter. 2007. Causation in Context. In Price and Corry, 191–223.

Nozick, Robert. 1969. Newcomb's Problem and Two Principles of Choice. In N. Rescher (ed.) *Essays in Honour of Carl G. Hempel*. Dordrecht: Reidel, 114–146.

Papineau, David. 2001. Evidentialism Reconsidered. *Noûs* 35: 239–259.

Paul, Laurie and Hall, Ned. 2013. *Causation: A User's Guide*. Oxford: Oxford University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Price, Huw. 1986. Against Causal Decision Theory. *Synthese* 67(2): 195–212.

———. 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 David Hull, Micky Forbes and Kathleen Okruhlik (eds.) *PSA 1992 Volume 2*. East Lansing, Michigan: Philosophy of Science Association, 253-267.

———. 2012. Causation, Chance and the Rational Significance of Supernatural Evidence *Philosophical Review* 121(4): 483–538.

Price, Huw and Corry, Richard (eds.). *Causation, Physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Price, Huw and Weslake, Brad. 2009. The Time-Asymmetry of Causation. In Helen Beebee, Christopher Hitchcock and Peter Menzies (eds.) *The Oxford Handbook of Causation*. Oxford*:* Oxford University Press, 414–44.

Russell, Bertrand. 1912–13. On the Notion of Cause. *Proceedings of the Aristotelian Society* New Series 13: 1–26.

Sobel, Jordan Howard. 1994. *Taking Chances: Essays on Rational Choice*. Cambridge: Cambridge University Press.

Skyrms, Brian. 1980. *Causal Necessity: A Pragmatic Investigation of the Necessity of Laws*. New Haven: Yale University Press.

Woodward, James. 2003. *Making Things Happen*. New York: Oxford University Press.

———. 2014. A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters—Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment). *Philosophy of Science* 81(5): 691–713.