View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Philsci-Archive

Causal Explanatory Power

Bejamin Eva[*] Reuben Stern[†]

August 11, 2017

Forthcoming in the British Journal for the Philosophy of Science[1]

Abstract

Schupbach and Sprenger (2011) introduce a novel probabilistic approach to measuring the explanatory power that a given explanans exerts over a corresponding explanandum. Though we are sympathetic to their general approach, we argue that it does not (without revision) adequately capture the way in which the causal explanatory power that c exerts on e varies with background knowledge. We then amend their approach so that it does capture this variance. Though our account of explanatory power is less ambitious than Schupbach and Sprenger's in the sense that it is limited to *causal* explanatory power, it is also more ambitious because we do not limit its domain to cases where c genuinely explains e. Instead, we claim that c causally explains e if and only if our account says that c explains e with some positive amount of causal explanatory power.

1 Introduction

There are many contexts in which we care how much the occurrence of an event causally explains the occurrence of another. Consider the recent presidential election in the United States. As we try to understand why President

[*] Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany) – http://be0367.wixsite.com/benevaphilosophy – benedgareva@icloud.com.

[†] Munich Center for Mathematical Philosophy, LMU Munich, 80539 Munich (Germany) – https://sites.google.com/site/reubenstern/research – reuben.stern@gmail.com.

[1] Both authors accept full and equal responsibility for what follows.

Trump won, we ask ourselves how much the election results are explained by Comey's timely announcement of the investigation against Hillary Clinton, and whether Comey's announcement better explains the election results than does Trump's promise to build a wall between the United States and Mexico.

Answering questions like these requires knowledge of the causal system at play. If we want to know how much c causally explains e then we must know whether and how c causes e.[2] But knowledge of the causal relationship(s) between c and e does not suffice for knowledge of the extent to which c causally explains e. This is because the causal influence that c exerts over e does not depend on what we know, while the extent to which c causally explains e does depend on what we know. Roughly put, the job of an explanation is to confer understanding, and since understanding can only be conferred to some agent(s), the extent to which c causally explains e should be specified relative to the understanding agent's (or agents') body of knowledge. The same is not true of the extent to which c causally promotes e since causal promotion is a worldly dependence relation that exists no matter whether there are any agents around to witness it.

Our task in this paper is to develop a method for assessing the extent to which c causally explains e given knowledge of the causal system at play. We follow in the footsteps of Schupbach and Sprenger (2011) in thinking that the explanatory power c has over e is (in some way) determined by the extent to which the occurrence of c makes the occurrence of e less surprising. However, we argue that Schupbach and Sprenger's own method for determining explanatory power is not sufficient for accurately capturing the way in which the explanatory power that c exerts over e depends on background knowledge in paradigmatic cases of causal explanation. Our task, then, is to develop a novel method of assessing explanatory power in terms of surprise reduction that fares better than Schupbach and Sprenger's approach when applied to causal explanations.

This means that our ambitions are a bit different from Schupbach and Sprenger's. They restrict the domain of their account to cases where c genuinely explains e, but do not restrict the domain of their account to

[2] Throughout the paper we adopt the convention of referring to variables with upper case letters and values of variables with lower case letters. Here, we make it clear that we are interested in the explanatory power that a *value* of a variable exerts over another.

cases where c causally explains e. Our account is less ambitious in the sense that it is limited to *causal* explanatory power, but more ambitious in the sense that we do not limit its domain to cases where c genuinely explains e.[3] Instead, we claim that c causally explains e if and only if our account says that c explains e with some positive amount of causal explanatory power.[4]

The outline of the paper is as follows. In section 2, we introduce Schupbach and Sprenger's measure of explanatory power. In section 3, we raise some issues surrounding the interpretation of the probability function used in the definition of their measure, before proposing an interpretation that bypasses these problems. Section 4 considers the problem of accounting for the influence of background knowledge on the evaluation of explanatory power. We show that standard Bayesian methods for accounting for such knowledge yield counterintuitive verdicts when applied to particular kinds of causal systems. Section 5 outlines our novel approach to quantifying causal explanatory power and shows how it naturally resolves the issues discussed in previous sections. Section 6 concludes and outlines new directions for future research.

Before going further, it is worth noting that this paper is situated within a broadly interventionist understanding of causation, of the type developed by e.g. Spirtes, Glymour and Scheines (2011), Pearl (2009), and Woodward (2005).[5] Although our approach makes extensive use of the causal modelling framework, we do our best to make the presentation self contained. Readers seeking a thorough introduction to the formalism should consult e.g. the first two chapters of Pearl (2009).

2 The Logic of Explanatory Power

In developing their approach to explanatory power, Schupbach and Sprenger (2011) start from the basic premise that

[3] Relatedly, our account applies only relative to a causal hypothesis, whereas their account can be applied *sans* any causal hypothesis.

[4] We discuss these issues in greater detail in Section 5.

[5] Of course, the term 'interventionist' is a vague one and there are important differences between the approaches forwarded by these authors. The sense in which our treatment of explanatory power is distinctively interventionist will become clear in later sections.

> [A] hypothesis offers a powerful explanation of a proposition...to the extent that it makes that proposition less surprising...for example, a geologist will accept a prehistoric earthquake as explanatory of certain observed deformations in layers of bedrock to the extent that deformations of that particular character, in that particular layer of bedrock, *et cetera* would be less surprising given the occurrence of such an earthquake.' (Schupbach and Sprenger 2011: 108).

This is a natural starting point that fits well with many everyday intuitions about explanation. When we try to explain the occurrence of an event,[6] e, we try to better understand why e's occurrence was rendered likely or necessitated by antecedent events. Consider President Trump's victory once more. As we try to explain what happened, we search for those events whose occurrence makes the election results unsurprising. It thus seems reasonable to say that there is at least a sense in which an event c explains e to the extent that c makes e less surprising or more expected.[7]

There may be multiple ways to formalise what it is for one event to make another less surprising. Schupbach and Sprenger work in the context of Bayesian probability theory. Specifically, one event c will render another event e less surprising if and only if $P(e) < P(e|c)$.[8] The strength of this inequality is called the 'statistical relevance' between e and c. In light of these considerations, Schupbach and Sprenger posit the following criterion as a basic requirement of any measure $\varepsilon(e, c)$ of the extent to which c explains e:

Positive Relevance (PR): Ceteris paribus, the greater the degree of

[6] Schupbach and Sprenger speak in terms of a hypothesis explaining some evidence, rather than in terms of an event explaining another event. We translate their discussion into event-speak because we find it more natural when discussing distinctively causal explanations. But it is worth mentioning that we are neutral with respect to whether a causal explanation of an event must include information about covering laws or explanatory generalisations (in addition to information about some antecedent event(s)). By our lights, and as will become clear later, the explanatory power that one event exerts over is accounted for by some covering laws and/or explanatory generalisations no matter whether such information must be included in the explanans. Woodward and Hitchcock (2003) present an account according to which such information must be included; Lewis (1986) argues that it needn't be.

[7] Section 5.4.2 considers competing senses in which an event can be rendered 'less surprising'.

[8] This assumes the standard definition of conditional probability as $P(e|c) = \frac{P(e \wedge c)}{P(c)}$.

statistical relevance between e and c, the higher $\varepsilon(e, c)$.

Philosophically, PR is the core idea behind Schupbach and Sprenger's approach to measuring explanatory power. It requires that, other things being equal, the extent to which c explains e should be a function primarily of the degree to which c raises the subjective probability of e. Broadly speaking, we are sympathetic to this intuition, and will use it as a guiding principle for the development of our measure of causal explanatory power. Crucially though, PR implies *ceteris paribus* that statistical relevance is both a necessary and a sufficient condition for an event c to have explanatory power over another event e.[9] Though this poses no problems for Schupbach and Sprenger (because they limit the domain to genuinely successful explanations), we will see that we must drop PR in order to distinguish genuinely successful causal explanations from unsuccessful causal explanations.

Schupbach and Sprenger posit six other criteria that $\varepsilon(e, c)$ is required to satisfy. Apart from PR, the only other criterion we need to consider requires that $\varepsilon(e, c)$ should be (partially) independent of the prior probability of c,

Irrelevance of Priors (IP): Values of $\varepsilon(e, c)$ do not (always) depend upon the values of $P(c)$.[10]

IP encodes the idea that the extent to which a given event c explains another event e is not dependent upon considerations of how likely c is in itself. For example, a tornado might be an excellent explanation of a broken weather vane, regardless of the fact that the prior probability of a tornado occurring is very low. We are measuring the extent to which c would render e less surprising, *in the event that c occurs.* Clearly, this quantity should be independent of our prior degree of belief in c occurring.

The seven criteria postulated by Schupbach and Sprenger are jointly sufficient to uniquely determine the following measure of explanatory power,

Definition 2.1 ***SSE:*** $\varepsilon(e, c) = \frac{P(c|e) - P(c|\neg e)}{P(c|e) + P(c|\neg e)}$

[9] The 'ceteris paribus' clause can be removed when we assume that c is genuinely explanatory of e.

[10] The condition that Schupbach and Sprenger actually use is slightly technical, but its central philosophical content is captured by IP as stated here.

For our purposes, the explicit form of $\varepsilon(e, c)$ will not be terribly important.[11] Our focus is on the fundamental assumption of PR and its implications. With Schupbach and Sprenger, we are happy to accept IP as a philosophical constraint on the notion of explanatory power.

3 Subjective and Nomic Distributions

3.1 Actual Degrees of Belief

Implicit in the Bayesian approach to explanatory power described above is the assumption that the probability functions deployed represent belief states. But whose belief states, and what kind of belief state? If the probability function is interpreted as reflecting the subjective belief state of the agent evaluating the explanation, then the Bayesian approach faces what we might call the 'explanatory old evidence problem'. For example, when we ask why Newcastle beat Sunderland, we usually know that Newcastle actually beat Sunderland, and therefore give the explanandum a subjective prior probability of 1.[12] But this is problematic. For e cannot be rendered less surprising by anything when it already has probability 1. And that's not all. The explanatory old evidence problem has an equally problematic flip side. As well as being certain that the explanandum occurred, we might also be certain about the occurrence of the prospective explanans. For example, when we are interested in how well the stock market crash of 1929 explains the Great Depression, we know that *both* of these events occurred, but c cannot raise the probability of e if c is already assigned probability 1. So in order to rescue the Bayesian approach, one must think of the probability function as expressing something other than our actual degrees of belief.

3.2 The Causal Distribution

Though the occurrence of the tornado may not make us more confident that the weather vane has broken (since we could have already seen both

[11] For detailed discussions of the virtues of this and other measures of explanatory power, see e.g. Cohen (2015), Crupi and Tentori (2012).

[12] Note that we are not committed here to the claim that explanation always presupposes certainty about the explanandum, only that it often does so.

the tornado and the broken weather vane), there is obviously a sense in which the tornado's occurrence makes it less surprising that the weather vane broke. What does this reduction of surprise consist in?

At first pass, it seems that an event c reduces the surprise of another event e to the extent that the occurrence of c provides reason to expect e given our best understanding of the causal system at play. That is, even when we know that there was a tornado and know that the weather vane is broken, the tornado's occurrence still makes the weather vane's breaking less surprising in the sense that our best understanding of the causal relationship between tornadoes and weather vanes tells us that we have strong reason to expect the weather vane to break in the event of a tornado.

This suggests that if we want to follow Schupbach and Sprenger in construing reduction of surprise as some kind of increase in probability, we should use probability distributions that encode what it is reasonable to expect about the events in question, given what we know about the causal system at play. Since these distributions must be consistent with the causal relationships (or structural equations) that govern the causal system, we refer to them as 'causal distributions'.[13]

For example, suppose that our best theory of the causal system governing the condition of the weather vane says that the probability of there being a tornado is .05 and that the probability of the weather vane breaking given that there is a tornado is .75, while the probability of the weather vane breaking given that there is not a tornado is only .005.[14][15]

[13] To be clear, we don't take ourselves to be defending a novel interpretation of probability. Rather, we simply mean to propose that the probabilities deployed are those that express the causal relationships in a causal hypothesis (no matter how they are interpreted). Thus there is still room to think of our approach as situated within a broadly Bayesian approach to explanatory power.

[14] For those who are familiar with structural equation models, the causal distribution for a causal system over a variable set V can be systematically recovered through specification, first, of the structural equations that describe the way in which the probability distribution over each variable depends on the probability distribution of its immediate causal predecessors or 'parents' and its error term, and, second, the unconditional probability distribution over all of the exogenous (parentless) variables. Since the probabilities over the exogenous variables represent reasonable prior expectations for these variables, and the structural equations represent the 'laws' that govern the causal system, the probability distribution that is implied by the structural equation model seems to give us exactly what we need.

[15] The reader may harbour doubts about our capacity to determine the unconditional proba-

The causal distribution includes these probability estimates, as well as every other probability estimate that is implied by the causal hypothesis at hand. Moreover, the causal distribution satisfies any constraints implied by whatever causal graph is under consideration. So if we consider a causal hypothesis according to which the only causal dependencies between X, Y, and Z can be represented as $X \rightarrow Y \leftarrow Z$, then X and Z must be probabilistically independent in the causal distribution.[16]

Since the causal distribution corresponds to what the causal details of the system give us reason to expect, rather than to our knowledge about what actually happens or has happened, the causal distribution is not obviously riddled with problems of old evidence. That is, even when we know that there actually was a tornado and that the weather vane actually broke, the causal distribution does not say that these events obtain with probability 1 because its probability estimates instead correspond to what the "laws" of the causal system tell us we have (or had) reason to expect.[17]

In some contexts, there may be reason to regard the laws of the causal system as true laws of nature–e.g., if we want to discuss whether c causally explains e relative to the (perhaps unknown) facts about the causal system. But we also might usefully deploy causal distributions that convey our presuppositions about the causal system at hand, rather than the truth about the causal system. For example, if we've just read a textbook about the causal relationship between c and e, then even if the textbook is wrong (unbeknownst to us), it is reasonable for us to assess whether and to what extent c causally explains e relative to the causal distribution that encodes what the textbook says about the causal system. In order to capture both of these senses, we do not require that the causal distribution is true. But we do suspect that there are contexts where it is reasonable to care exclusively

bility estimates of exogenous variables since these are not supplied by the structural equations. We are sympathetic to these concerns and are correspondingly ecumenical about how these probabilities should be set. In the absence of knowledge of long-run frequencies, one might justifiably use tools in the objective Bayesian toolbox to determine these estimates–e.g., principles of indifference.

[16] This follows from the Causal Markov Condition (see section 4).

[17] The causal distribution assigns probability 1 to an event only if the causal hypothesis at play says that the event will certainly occur. Although we associate these probabilities with the 'laws', we do not take this to constitute a necessary deviation from standard Bayesian interpretations of probability.

about whether c explains e relative to the truth, as well as contexts where it is reasonable to care exclusively about whether c explains e relative to our presuppositions (regardless of their truth).

No matter whether the causal distribution must be true, it is clear that the notion of surprise that underwrites our treatment of causal explanatory power is somewhat distinct from the notion that is interpreted in terms of actual degrees of belief. Since the strength of c as an explanation for e depends on the extent to which the occurrence of c renders e less surprising according to reasonable (causal) expectation, rather than the extent to which c renders e less surprising to actual agent(s), the explanatory power that c exerts over e is better understood as depending on the extent to which c would have rendered e less surprising (given what we know about the causal system) than as depending on the extent to which c actually does render e less surprising.[18]

Alhough opting for causal distributions allows us to avoid issues that arise when c and e are known, it gives rise to new problems because there are certain pieces of background knowledge (not pertaining to c and e themselves) that seem to affect explanatory power. In the next section, we begin to develop an account of explanatory power that can accommodate background knowledge as needed.

4 Background Knowledge

Consider the causal system in figure 1 (based on examples from Hesslow (1976)). Birth control and pregnancy both cause thrombosis. However, birth control causally inhibits pregnancy. It is possible for this system to be 'unfaithful',[19] where BC and TH are statistically independent, regardless of

[18] It is tempting to say that the extent to which c explains e does depends on the extent to which c would render e less surprising had c not happened, but the phrase 'had c not happened' is ambiguous between 'had c not happened with certainty' and 'had c occurred with the non-extreme probability suggested by the causal system'. If 'had c not happened' is understood in the latter way, we are happy with this characterization of the counterfactual query. But our survey of the literature suggests that most philosophers understand 'had c not happened' in the former way (see e.g. Briggs 2012, Halpern 2000, Woodward 2005).

[19] When we say that a system is unfaithful, we mean that the assumed probability distribution/graph pair does not satisfy the Causal Faithfulness Condition (see Spirtes, Glymour and

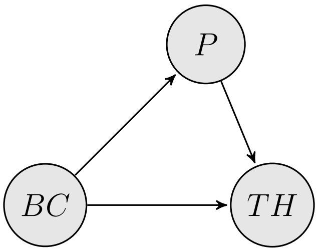

Figure 1: Birth Control, Pregnancy, Thrombosis

the fact that the causal model represents TH as being causally dependent on BC. Intuitively, this will happen when the two paths from BC to TH 'cancel out' exactly. It might be that, although birth control causally promotes thrombosis through some physical mechanism, this effect is completely and perfectly negated in a statistical sense by the fact that birth control is negatively correlated with pregnancy, which is itself a cause of thrombosis. In such a case there will be no correlation between birth control and thrombosis, despite the causal connection between the two phenomena. It is easy to see that $\varepsilon(th, bc) = 0$ will then be guaranteed to hold. So our attempted explanation of Suzy's thrombosis by her taking birth control pills will have no power whatsoever, which seems correct. If we know nothing about whether or not Suzy has been pregnant, then her having taken birth control would not render her thrombosis any less surprising (since taking birth control does not raise the probability of getting thrombosis in the causal distribution).

But now suppose that we also happen to know that Suzy has never been pregnant. In this case, it seems clear that Suzy's taking of birth control pills very strongly explains her thrombosis. For in this case our knowledge of Suzy's pregnancy breaks the correlation between pregnancy and birth control, thereby inducing a strong correlation between birth control and thrombosis. Given that we know that she is not pregnant, Suzy's taking of birth control pills renders her thrombosis much less surprising.

Thus, we should think that bc exerts considerable explanatory power over th. However, the causal interpretation of the probability distribution prevents us from straightforwardly obtaining this result. This follows from

Scheines (2000)). The Causal Faithfulness Condition is satisfied by a given graph/probability distribution pair exactly when all *and only* the independencies entailed by the Causal Markov Condition (discussed in detail later in this section) obtain.

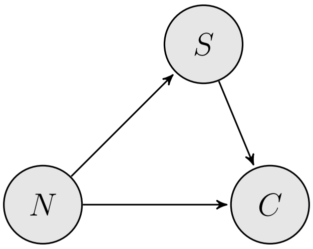

Figure 2: Smoking, Cancer, Nicorette

the fact that the causal probability of pregnancy does not correspond to our subjective degree of belief about the particular instantiation of that causal system. We need some way to update the causal distribution to allow it to take into account our background knowledge which, as we have seen, can have a significant effect on our judgements of explanatory strength.

Perhaps it seems like this case can be bracketed since it involves the cancellation of causal paths, and such cancellation rarely occurs. But this is not right. Consider the causal system in figure 2, according to which chewing Nicorette reduces the probability that one subsequently smokes (and thereby significantly reduces the probability of the later onset of cancer), but also increases the probability of cancer along a distinct path (because of the deleterious effect that chewing Nicorette has on the body).

Here, we can easily imagine that Nicorette's total effect is healthy in the sense that it reduces the probability of cancer because the causal path that is mediated by smoking is stronger than the other.[20] Thus if we know nothing about whether John stops smoking after chewing Nicorette, his taking Nicorette appears to increase the surprise (by reducing the causal probability) of his cancer. But if we know that John quits smoking (upon chewing the Nicorette), but does subsequently get cancer, it seems that Nicorette does causally explain his cancer to some extent. Thus, just as the introduction of background knowledge can render c explanatorily relevant to e when it otherwise wouldn't be (as in Suzy's case), it also can change the extent to which c is explanatorily relevant to e in cases where c is explanatorily relevant to e sans background knowledge (as in John's case).

[20] See Woodward (2003) for helpful discussion of total effects, contributing effects, and direct effects.

Before proposing a solution to this problem, it is useful to make two small clarifications about our general approach. Firstly, the account presented here applies most naturally to *token-level* causal explanations, as opposed to *type-level* explanations. In the previous case, for example, it seems that we are interested not in the question of how much the taking of birth control explains thrombosis in general, but rather in how much particular instances of thrombosis are explained by particular instances of the taking of birth control pills. When considering questions of this kind, it is clear that our judgements typically depend on our background knowledge of the facts surrounding the relevant instantiation of the causal system.

Secondly, philosophers sometimes speak in terms of 'how-possibly' and 'how-actually' explanations. We are interested in both. That is, we want a treatment that works well both when we are trying to explain why e *did* occur and also why e *might* occur.

At this juncture, it is also useful to introduce Pearl's (2009) language of 'd-separation' in order to facilitate discussion of the the way in which causal explanatory power varies with background knowledge. The language of d-separation is useful for characterising the probabilistic conditional independencies that are entailed by the Causal Markov Condition (an axiom of the graphical approach to causal modeling), given a directed acylic graph (DAG) whose edges depict the causal relations at play.[21] The Causal Markov Condition is the assumption that encodes the fact that causes "screen off" their effects and is likewise presumed by the interventionist approach to causation (see e.g. Hausman and Woodward (1999)). If every (undirected) path between a pair of variables is d-separated by Z (according to a given causal graph), then the pair of variables must be probabilistically independent of each other conditional on any assignment of values over Z, where d-separation is defined as follows,

d-separation: A path between two variables, X and Y, is d-separated (or blocked) by a (possibly empty) set of variables, Z, if and only if

i the path between X and Y contains a non-collider that is in Z, or

ii the path contains a collider, and neither the collider nor any descendant

[21] A DAG graphically represents the causal relations that obtain among a set of variables V as a set of directed edges (or arrows), such that no directed path forms a cycle.

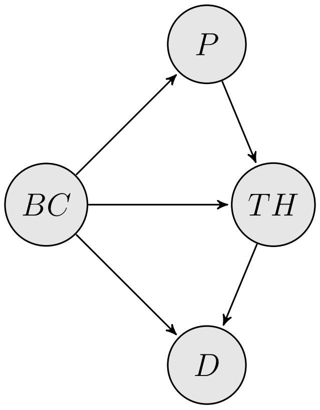

Figure 3: Birth Control, Pregnancy, Thrombosis, Death

of the collider is in Z.

Collider: A variable is a collider along a path if and only if it is the direct effect of two variables along the path. (So M is a collider along $I \to M \leftarrow J$ but not along $I \leftarrow M \to J$ or $I \to M \to J$.)

4.1 Conditionalisation and Colliders

Returning to the problem at hand, there is an obvious fix available here. In particular, we can simply condition on our background knowledge to update the causal distribution. So, in Suzy's case, we continue to use the causal distribution but simply condition on the knowledge that Suzy is not pregnant. Since P d-separates BC and TH along the top path and bc is strong evidence for th conditional on either value of P, $\varepsilon(th, bc)$ is forced to have a high value. However, this solution quickly breaks down when we consider slightly more complex causal systems.

Suppose we also represent whether Suzy dies with another variable D (see figure 3). Plausibly, death is causally influenced by both the taking of birth control pills and thrombosis.[22] If we have no knowledge about whether Suzy is pregnant, then thrombosis is independent of birth control, and this rules out birth control as a viable explanation. But now suppose that we happen to know that Suzy tragically went on to die. Following the

[22] For instance, imagine that birth control pills have as a positive side-effect that they causally inhibit the occurence of particular types of cancer, and so are negatively correlated with death.

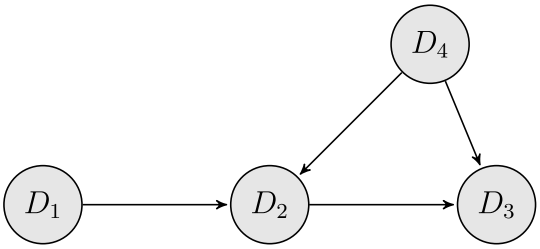

Figure 4: Dominoes 1-4

procedure of conditioning on background knowledge, we should condition on our knowledge that Suzy died. Unfortunately, this leads to some dubious results. Crucially, death is causally downstream of both birth control and thrombosis. So updating our knowledge of D involves conditioning on a collider between BC and TH. This means that we may introduce a spurious correlation between birth control and thrombosis. If birth control has a negative direct causal effect on death (which is conceivable) then the nature of this new correlation will be positive. This will mean that $\varepsilon(th, bc)$ will be positive after we condition on $D = d$, whereas it was zero before we did so, i.e. conditioning on death gave the explanation of thrombosis by birth control power where it previously had none. But this seems wrong. D is causally downstream of both the explanans and the explanandum, and our knowledge of its value should have no bearing on our assessment of the prospective causal explanation. Indeed, if one thinks only about temporal order then this point become even more apparent. How can death affect the extent to which birth control causally explains thrombosis when it only occurred after the birth control pills had been taken and the subject had developed thrombosis? Clearly, something is wrong here.

Another amendment suggests itself. In particular, it is tempting to impose the condition that we only update on knowledge that is not causally downstream of the explanandum. This deals nicely with the case discussed above without requiring any major ammendments to the overall strategy. Sadly though, it doesn't work. To see this, consider the causal system in figure 4.

Here, we are considering a system of four dominoes $(D_1, ..., D_4)$. D_1, D_2, D_3 are lined up in a row and D_4 is placed off to the side, equidistant from D_2 and D_3. The nodes in the network each have two possible values, i.e. the dominoe either does or doesn't fall. D_1 has some prior probability of falling, and if it falls, it will knock down D_2 with some probability, and D_2 will then have some probability of knocking down D_3. Similarly, D_4 has some prior probability of falling, and once it falls there is some probability that it knocks down D_2/D_3.

We are interested in knowing how well D_1 falling explains D_3 falling. Suppose that we know that D_2 fell. In this case, when explanatory power is understood in terms of reduction of surprise, it seems like D_1 falling should exert zero explanatory power over D_3 falling, since D_1 can only influence D_3 via D_2.[23,24] So if we know that D_2 fell, then any knowledge about D_1 should be explanatorily irrelevant to D_3. Once we know that D_2 fell, D_1 falling doesn't tell us anything new about *why* D_3 fell.

Now, since D_2 is not causally downstream of the explanandum (D_3) we are allowed to condition on this knowledge. But since D_1 and D_4 are d-connected by D_2, it is consistent with the Causal Markov Condition that there is a correlation between the falling of the first and fourth dominoes when we condition, even though there is no explanatory power. Thus, by conditioning on our background knowledge, we get the wrong result again. And this time we can't blame the problem on the fact that the background knowledge is downstream of the explanandum. There is something else going wrong at a deeper level.

4.2 A Helpful Intervention

Even with the restriction that we only update on background knowledge that is not downstream of the explanandum, conditioning on background knowledge appears to affect our judgments of explanatory power in an unintuitive way. Luckily, there is an alternative approach available to us. In particular, we can invoke the notion of a counterfactual 'intervention'. The proposal then is to *intervene* on the relevant items of background knowledge, rather than conditioning.

[23] See section 5.4.1 for further discussion of possible objections that may arise for this sort of case.

[24] Contrast this with the case where we know nothing about whether D_2 fell. In that case, D_1's falling *does* look like a very good explanation of D_3's falling, since it does seem to tell us a lot about why D_3 fell.

One can compute the effect of intervening on a causal system to set a variable to a particular value. Following Pearl (2009), Spirtes, Glymour and Scheines (2000), and Woodward (2005), allow the intervention on X to represent some justifiably omitted cause of X that can be exploited to set X to a value x. Since omitted causes cannot be common causes, the intervention on X must be d-separated and therefore independent from all of X's non-descendants.

Consider again the causal system depicted in figure 4. As before, suppose that we happen to know that D_2 fell. In this case, D_1's falling should exert *no* explanatory power over D_3 falling. This is because all of D_1's potential explanatory power is via D_2, and we already know that D_2 fell (the causal influence that D_1 exerts over D_3 all travels through D_2). We saw that simply conditioning on our knowledge that D_2 fell violates these intuitions. But according to our new proposal, we should rather *intervene* on the causal system to ensure that D_2 falls. Since the intervention to make D_2 fall is d-separated from D_1 (and therefore not correlated with D_1), intervening effectively 'breaks' the edge between D_1 and D_2. (This may be intuitive since there is a sense in which exogenous interventions are disruptions to the causal system.) This allows us to set the probability of d_2 to 1 (accurately reflecting our background knowledge), but because we have broken the edge between D_1 and D_2, no new correlations between D_1 and D_3 via D_4 are introduced. Thus we get the good without the bad. D_1 and D_3 are now statistically independent, meaning that $\varepsilon(d_3, d_1) = 0$ will have to hold, as desired. Our proposal naturally provides the right answer in this case.

Furthermore, our new proposal makes the artificial condition that we should only take into account background knowledge that is not downstream of the explanandum unneccessary. Consider again the system in figure 3 and suppose that we know that Suzy went on to die, but we have no other background knowledge about her particular case. According to the previous proposal, we should condition on Suzy's death, but as we saw this affects our evaluation of the explanatory relationship between bc and th in a problematic way. Following the new procedure, we propose that we should rather intervene to ensure Suzy's death.[25] Because the intervention on D is d-separated from the causal predecessors of D (unlike D itself), the correlation between between BC and TH remains unaffected, so $\varepsilon(th, bc) = 0$

[25] No Suzies were harmed in the writing of this paper.

continues to hold, as desired.[26] Thus, the new procedure is equally able to deal with background knowledge of events that are causally downstream of the explanandum. This means that we no longer need to impose by hand the artificial restriction that some of our background knowledge about the particular case should be ignored when we're evaluating explanatory strength.

In light of these examples, we conclude that when updating the causal distribution to take into account the background knowledge of the agent assessing the strength of an explanation, we should always *intervene* (and not condition) on the relevant background knowledge. More generally, the examples discussed so far clearly illustrate the importance of taking into account background knowledge when assessing the strength of a prospective causal explanation.

5 Causal Explanatory Power

Let's summarise our progress so far. We saw that the Bayesian faces two old-evidence style problems when probabilities are interpreted as actual degrees of belief. Happily, these problems were resolved by opting for the causal distribution. However, this led to a new problem regarding how to take into account the background knowledge of the agent evaluating the explanation. We saw that taking an interventionist approach to background knowledge gave intuitively correct answers in cases where standard Bayesian conditionalisation failed. So now that we have a satisfactory interpretation of the relevant probability distributions and are able to account for the context sensitivity of causal explanatory power, all that remains is to find a suitable measure. At this stage, it is tempting to simply adopt the following procedure.

(1) First, represent the causal system in which the explanans and explanandum are embedded and work out the causal distribution for that system, according to our current best knowledge.

[26] Woodward (2005) famously exploits this feature of interventions when definining 'direct cause' and 'contributing cause'. By intervening on colliders, we do not induce spurious associations that would arise were we to simply condition.

(2) Second, update on *all* your background knowledge (excluding only the explanans and the explanandum themselves) regarding the causal system by intervening to set the relevant variables to their known values.

(3) Calculate the explanatory power that the explanans exerts over the explanandum using the measure ε, relative to the updated causal distribution.

On this approach, we would continue to use Schupbach and Sprenger's measure ε, but it would now be measuring statistical relevance relative to a causal distribution that has been updated via interventions on the relevant items of background knowledge.

5.1 The Applicability of Explanatory Power

Before going further, it makes sense to take into account a significant caveat that Schupbach and Sprenger place on the applicability of ε. Specifically, they write

> [W]e restrict ourselves ... to speaking of theories that do in fact provide explanations of the explanandum in question. This account thus is not intended to reveal the conditions under which a theory is explanatory of some proposition (that is, after all, the aim of an account of explanation rather than an account of explanatory power); rather, its goal is to reveal, for any theory already known to provide such an explanation, just how strong that explanation is. (Schupbach and Sprenger 2011: 107)

To translate this passage into causal language, the measure ε is only supposed to be applied to genuine causal explanations. We cannot apply the measure in cases where the explanans has no causal explanatory bearing whatsoever on the explanandum. Rather, we should rely on some external theory of causal explanation to delineate for us the class of genuine causal explanations, to which we can subsequently apply the measure. The measure itself cannot decide for us whether or not one event causally explains another, but given that such an explanatory relationship exists, it can tell us the strength of that relationship.

And this is fine, as far as it goes. However, it would be nice if our measure of causal explanatory power was itself capable of determining, for

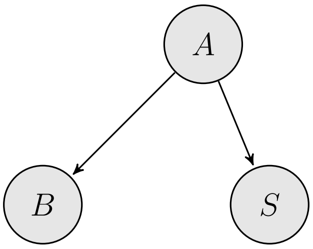

Figure 5: Storm, Barometer, Atmosphere

any explanandum event e and any other event c, whether or not c is causally explanatory of e. To see why this would be desirable, consider an analogy. It might be that we are only interested in using scales to measure the weight of things with mass, but it would also be nice if the scales read zero whenever there is no mass to weigh.

5.2 Statistical Relevance $\neq$ Causal Explanatory Power

To see why ε can only be applied to genuine causal explanations, consider the example depicted in figure 5. A is a binary variable corresponding to the presence (or absence) of a sudden drop in atmospheric pressure. S and B represent the coming of a storm and a drop in the Barometer, respectively. Now, suppose that we want to explain a storm by a drop in the Barometer. Clearly, this is a terrible causal explanation that should have zero strength. However, B and S are both highly correlated with A. Since B and S are d-connected and we have no background knowledge about A, B could have a high degree of statistical relevance to S. It is not hard to see that $\varepsilon(s, b)$ will generally be high in such a setup, which conflicts with our very strong intuitions that the drop in the barometer should have no causal explanatory strength with respect to the storm. The upshot of this example is that statistical relevance is not sufficient for causal explanatory strength, and so ε, as a measure of statistical relevance, is bound to give the wrong answer in such cases.

As noted above, Schupbach and Sprenger actually anticipate this sort of problem. Since b does not provide a genuine causal explanation of s, the measure simply shouldn't be applied in this kind of case. However, it is independently desirable to have a measure that is itself able to detect when an event c fails to provide a causal explanation of an explanandum e. We now turn our attention to providing such a measure.

5.3 Interventionist Explanatory Power

Finally, we propose the following procedure for computing explanatory power,

(1) First, represent the causal system in which the explanans and explanandum are embedded and work out the causal distribution for that system, according to our current best knowledge.

(2) Second, update on *all* your background knowledge (excluding only the explanans and the explanandum themselves) regarding the causal system by intervening to set the relevant variables to their known values.

(3) Calculate the explanatory power that *intervening* to make the explanans true exerts over the explanandum using the measure ε, relative to the updated causal distribution.

Thus, according to the new procedure, the explanatory power that an explanans c exerts over an explanandum e will be defined as $\mathfrak{E}(e, c) = \varepsilon(e, do(c))$, where ε is Schupbach and Sprenger's measure of statistical relevance and $do(c)$ represents our intervention to make c occur.[27]

[27] One might be concerned that the new procedure for calculating explanatory power suffers from a basic flaw. Suppose, for example, we want to explain the storm by the drop in the barometer and we happen to know that there was a drop in atmospheric pressure. Following our procedure, we (i) adopt the causal distribution for the system, (ii) update on our background knowledge by intervening to make $A = a$ true, (iii) measure the relevance of $do(B = b)$ to $S = s$, i.e. intervene to make $B = b$ true and see how the probability of $S = s$ changes in the updated causal distribution. But now, suppose that we are also interested in evaluating the explanatory power of a drop in atmospheric pressure for the presence of a storm, and we happen to know that the barometer dropped. Following the procedure, we would (i) adopt the causal distribution, (ii) update on our background knowledge by intervening to make $B = b$ true, (iii) measure the relevance of $do(A = a)$ to $S = s$, i.e. intervene to make $A = a$ true and see how the probability of $S = s$ changes in the updated causal distribution. Intuitively, these two explanations should receive very different scores. The first one seems like a very bad explanation while the second one looks like a potentially good explanation. But, it may seem that $\mathfrak{E}$ will have to assign them equal explanatory power because in both cases we end up intervening on $B = b$ and $A = a$

To justify this new procedure, consider again the example from figure 5. If we have no relevant background knowledge, then P is just the standard causal distribution. As before, suppose that we attempt to explain the storm by the dropping of the barometer. It is not hard to see that intervening to make the barometer drop is independent of the presence of a storm. Specifically, intervening to make b true will sever the causal link between A and B, thereby rendering S statistically independent of B. So $P(s|do(b)) - P(s) = 0$, i.e. $\mathfrak{E}(b, s) = 0$. And this is exactly the result that we wanted. The dropping of the barometer has no explanatory power with respect to the storm. Simply applying the measure ε to the explanandum and the explanans gives the wrong result here, but applying ε to the explanandum and the *intervention* on the explanans gives the right answer.[28]

Thus, the new measure $\mathfrak{E}$ has a much broader domain of application than Schupbach and Sprenger's measure. $\mathfrak{E}$ doesn't rely on any external theory of explanation in order to determine the candidate explanations to which it can apply. Rather, $\mathfrak{E}$ can be applied to *any* prospective causal explanation and will itself determine whether or not it really deserves the name, and if so, to what extent.

Although this new approach departs significantly from the original statistical relevance criterion PR, it still respects the intuition that a strong explanation is one that makes the explanandum less surprising, but in a

and measuring how the probability of $S = s$ changes. The only difference is the order in which we intervene on B and A, and since interventions commute (as long as they are on different variables), it may seem that this will make no difference (see Redacted).

Luckily, this objection rests on a basic misunderstanding. To see this, let's introduce some new terminology. Let C be the causal distribution for the system, and let C_U be the result of updating the causal distribution according to our background knowledge. Clearly, C_U will be a different function in each case (Since we have updated on different pieces of background knowledge in the two cases) and so the degree to which s is subsequently confirmed by intervening on the candidate explanandum will differ between the cases. The confusion stems from the fact that the overall degrees of confirmation $C(s|do(b), do(a))$ and $C(s|do(a), do(b))$ are the same, but this does not imply that $\mathfrak{E}(s, b)$ and $\mathfrak{E}(s, a)$ will be equal. And indeed, it is not hard to see that our approach will provide the correct result that $\mathfrak{E}(s, a)$ will be much greater than $\mathfrak{E}(s, b)$ in these cases.

[28] It should be noted that since $\mathfrak{E}$ is defined as applying only to $do(c)$, $\mathfrak{E}$ fails to apply to disjunctive explanations, since it is not clear how to intervene to make a disjunction true. A similar issue arises with conjunctions (but it has been suggested that this can be solved through the use of multiple interventions).

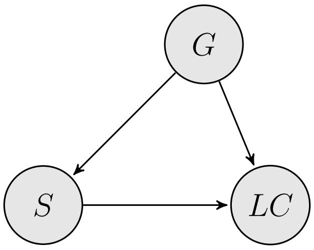

Figure 6: Smoking, Lung Cancer, Gene

different sense. On our approach, c is a good causal explanation for e to the extent that intervening to make c occur would render e less surprising in the causal distribution. And this seems right. By intervening to make c true, we discard all non-causal correlations between c and e and isolate the genuinely causal relationships between c and e, which are all we should take into account when assessing the causal explanatory power of c for e.

5.4 $\mathfrak{E}$ Illustrated

Consider the example depicted in figure 6. We suppose that there is a gene that (i) makes people more likely to smoke, (ii) makes people less likely to develop lung cancer. As usual, suppose that smoking also causes lung cancer. Suppose finally that this is an unfaithful causal system, i.e. smoking and lung cancer are statistically independent even though smoking causally promotes lung cancer. This is possible since smoking is now positively correlated with another causal inhibitor of lung cancer ($G = g$). However, smoking could still be a very good explanation for somebody developing lung cancer. So we have significant explanatory power and no statistical relevance. Statistical relevance is neither necessary nor sufficient for explanatory power. Contrast this with Hesslow's example in figure 1. In that case, the explanans (bc) is also statistically independent of the explanandum (th). However, in that case, taking birth control does not *make* thrombosis any less probable (because the canceling paths are causal), so it seems intuitive that there should be zero explanatory power. Equating explanatory power with statistical relevance gives the result that smoking has no explanatory power for lung cancer and birth control has no explanatory power for thrombosis. We only want the second of these results to obtain.

Happily, the new measure $\mathfrak{E}$ gets these cases exactly right. In the case from figure 6, we intervene to give S a positive value. In doing so we break the edge between S and G and leave S positively correlated with LC. The net effect is that lc gets confirmed, meaning that $\mathfrak{E}(lc, s)$ will have a positive value, as desired. In the case from figure 1, we intervene to give BC a positive value. But since BC has no parents, this is just the same as conditioning on $BC = bc$ and since BC is independent of TH, th doesn't get confirmed. So $\mathfrak{E}(th, bc) = 0$ will hold, as desired.

Another important question concerns the relationship between $\mathfrak{E}$ and Schupbach and Sprenger's condition IP, which requires that the explanatory power of c for e is not generally proportional to the prior probability of c. It is easy to see that $\mathfrak{E}$ will naturally satisfy IP. For, $\mathfrak{E}(e, c)$ measures the extent to which intervening to make c occur is statistically relevant to e's occurence, and this is obviously not proportional to the prior probability of c. Indeed, the opposite will often be the case. If c has an extremely high prior probability, then intervening to make c occur generally won't have much of an effect on the probability of e. So, ceteris paribus, highly likely causes have low explanatory power for their effects. This suggests that our approach to explanatory power may naturally capture our preference for 'abnormal causes' (see e.g. Halpern and Hitchcock (2014)).[29][30]

[29] One prominent issue in the causal explanation literature concerns the phenomenon of causal overdetermination. To illustrate, suppose that there are two extremely accurate marksmen M_1 and M_2, and a potential gunshot victim who might die, D. Further, suppose that whether M_1 shoots is independent of whether M_2 shoots, but if either of them shoot, D is extremely likely. Now, suppose that we want to explain D by M_1 shooting. If we happen to know nothing about whether or not M_2 shot, then our account gives the intuitively correct verdict that this is a very good explanation, since M_1's shooting renders D much less surprising. However, if we also happen to know that M_2 shot, then M_1 shooting would be a very poor explanation since intervening to make M_1 shoot would not make D significantly less surprising (because D is already near-guaranteed by M_2's shooting). Thus, our approach to causal explanatory power appears to give intuitive results for cases of causal overdetermination. This is a topic we will explore in more detail in future work.

[30] At this juncture, it is also worth mentioning that, despite some surface similarity, our account of causal explanatory power is very different from the accounts proposed by Halpern (2016). The essential difference is that according to any of the notions of explanatory power suggested by Halpern, the extent to which c explains e does not track the extent to which c renders e less surprising. Our focus on surprise reduction is what allows our approach to favour abnormal causes without any additional apparatus.

5.4.1 An Objection

Finally, consider the following example. Ettie's Dad went to see the local football team play in a crucial end of season match. Unfortunately, Ettie was busy on the day of the game, so she couldn't go with him. On her way home, she read a newspaper headline saying that the local team had lost. When she got home, she asked him 'Dad, why did we lose?', to which her witty father replied 'because we were losing by fifty points when the fourth quarter started'. Understandably, Ettie still wanted to better understand why her team lost, so she asked her Dad why they were down by so much entering the fourth quarter. He replied that their best player was injured in the opening minutes of the game, and, finally, Ettie's curiosity ran out.

This may look like a problem for our account in a couple ways. First, it may seem that when Ettie's Dad answered her initial question, a better answer would have been that the best player got injured, even though $\mathfrak{E}$ will assign his actual answer a much higher degree of explanatory power (since a massive fourth quarter deficit renders the loss more certain than does a first quarter injury). Second, as the story is told, Ettie asks her second question in order to better understand why her favorite team lost, but according to $\mathfrak{E}$, the player's injury does not qualify as explanatory of the loss (since her background knowledge of the fourth quarter deficit would seem to screen off the injury from the loss).

Regarding the first problem, there are a couple of things to say. Firstly, this problem is not unique to our approach. Schupbach and Sprenger's measure ε will be equally susceptible to such examples, since trailing by fifty points in the fourth quarter is more highly correlated with losing than is an early injury to one's best player, and ε tracks statistical relevance. Indeed, any approach that tracks the reduction of surprise will treat this case in this way. Secondly, and more importantly, there may be more than one sense in which one explanation can be more successful than another. On the one hand, being down by fifty points entering the fourth very strongly causally explains the loss (in much the same way that the penultimate domino's falling strongly explains the last domino's fall) by almost entirely eliminating any surprise. On the other hand, the injury may be a *deeper* explanation (perhaps because it explains *more* facts about the causal system). We believe that explanatory depth is worthy of philosophical investigation, but the measure $\mathfrak{E}$ should be taken as a measure of explanatory *power*, not explanatory depth.[31] Moreover, relative to the empty set of background knowledge, our account does capture the sense in which the injury explains *more*, since the injury is explanatorily relevant to two variables in the system, while fourth quarter deficit is explanatorily relevant to only one.

Regarding the second problem, when we think carefully about this case, upon hearing her Dad's initial answer, Ettie acquires a new fact that she wants explained–namely, the massive fourth quarter deficit. Upon doing so, it seems that Ettie's explanatory target *shifts*, and that the player's injury is (relative to her background knowledge) a strong explanation of the new explanatory target. Our account gets this exactly right.

6 Conclusion

In summary, we have presented a novel method for determining the extent to which an event c causally explains another event e. Unlike others before us, our approach successfully captures the effect that background knowledge has on judgements of explanatory power. Moreover, our account can distinguish between genuine and illegitimate causal explanations without recourse to any external theory of explanation.

In future work, we hope to explore

i How and whether these resources can be used to understand and refine the notion of 'explanatory depth'.

ii The bearing of this approach upon questions surrounding our preference for abnormal causes.

iii How and whether our account can be extended to deal with'multi-level' explanations in which synchronic dependence relations such as supervenience play a role.

iv The possibility of empirically testing our account of causal explanatory power as predictive of the way people actually make explanatory judgements.

[31] See Hitchcock and Woodward (2003) for a good discussion of explanatory depth.

References

Briggs, R. (2012). *Interventionist Counterfactuals. Philosophical Studies*, 160: 139-166

Cohen, M. (2016). *On Three Measures of Explanatory Power with Axiomatic Representations. British Journal for the Philosophy of Science*, 67 (4): 1077-1089

Crupi, V and Tentori, K. (2012). *A second look at the logic of explanatory power (with two novel representation theorems). Philosophy of Science*, 79: 365-285

Halpern, J. (2000). *Axiomatizing Causal Reasoning. Journal ofArtificial Intelligence Research*, 12: 317-337.

Halpern, J. and Pearl, J. (2001). *Causes and Explanations: A Structural-Model Approach- Part 2: Explanations.* In *Proceedings of the Seventh Joint Conference on Artificial Intelligence*, San Francisco: Morgan Kaufmann.

Halpern, J. (2016). *Actual Causality.* MIT University Press: Cambridge.

Halpern, J. and Hitchcock, C. (2014). *Graded Causation and Defaults. British Journal for the Philosophy of Science*, 66(2): 413-457.

Hausman, D. and Woodward, J. (1999). *Independence, Invariance and the Causal Markov Condition. British Journal for the Philosophy of Science*, 50: 521-583.

Hitchcock, C and Woodward, J. (2003). *Explanatory Generalizations, Part 2: Plumbing Explanatory Depth. Nous*, 37(2): 181-199.

Hesslow, G. (1976). *Discussion: Two Notes on the Probabilistic Approach to Causality. Philosophy of Science*, 43: 290-292.

Lewis, D. (1986): *Causal Explanation.* In *Philosophical Papers: Volume 2*: 214-240. Oxford: Oxford University Press

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

Schupbach, J. and Sprenger, J. (2011). *The Logic of Explanatory Power. Philosophy of Science* 78(1): 105-127

Spirtes, P., Glymour, C., and Scheines, R. (2000): *Causation, Prediction and Search*, Cambridge, MA: MIT Press.

Woodward, J. and Hitchcock, C. (2003). *Explanatory Generalizations, Part 1: A Counterfactual Approach. Nous*, 37(1): 1-24.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*, Oxford Studies in the Philosophy of Science. Oxford: Oxford University Press.