CAUSAL HISTORY, STATISTICAL RELEVANCE, AND EXPLANATORY POWER

DAVID KINNEY

This paper is forthcoming in a supplementary issue of *Philosophy of Science* containing selected papers from PSA 2022. Final contents are subject to potential minor revision. Please cite the published version when it is available. I am grateful to Jonathan Birch, Luc Bovens, Christopher Hitchcock, Christian List, Katie Steele, Reuben Stern, Thalia Vrantsidis, and audiences at the LSE PhD student work-in-progress seminar, the 2018 Explanatory Power Workshop at the University of Geneva, and the Concepts and Cognition Lab at Princeton University for feedback on various drafts of this paper.

Abstract. In discussions of the power of causal explanations, one often finds a commitment to two premises. The first is that, all else being equal, a causal explanation is powerful to the extent that it cites the full causal history of why the effect occurred. The second is that, all else being equal, causal explanations are powerful to the extent that the occurrence of a cause allows us to predict the occurrence of its effect. This article proves a representation theorem showing that there is a unique family of functions measuring a causal explanation's power that satisfies these two premises.

1. Introduction

Several authors in philosophy of science have argued that, all else being equal, a causal explanation is good to the extent that it provides a detailed description of the causal history of why the event being explained (i.e., the *explanandum*) occurred. Consider the example from Railton (1981):

> For any given gas, its particular state S at a time t will be determined solely by its molecular constitution, its initial condition, the deterministic laws of classical dynamics operating upon this initial condition, and the boundary conditions to which it has been subject. Therefore, the ideal explanatory text for its being in state S at time t [...] will be a complete causal history of the time evolution of that gas (p. 250).

The idea being expressed here is that the ideal causal explanation of why a gas ends up in state S at a time t is the full causal history of the gas' evolution from some state S', at some previous time t', to its state S at t. From this exemplar of an *ideal* causal explanation, one can make the further inference that causal explanations in general are good or powerful to the extent that they approximate this ideal. One finds a similar idea expressed by Salmon (1984), who holds that in many cases, good explanation "involves the placing of the explanandum in a causal network consisting of relevant causal interactions that occurred previously and suitable causal processes that connect them to the fact-to-be-explained" (p. 269). Similarly, Lewis (1986) defends the thesis that "to explain an event is to provide some information about its causal history" (p. 217). Keas (2018) also defends the idea that, all else being equal, scientific explanations are good to the extent that they trace the causal history of an event back as far as possible, calling this feature of an explanation "causal history depth."

On the other hand, there is also widespread agreement in the literature that, all else being equal, explanations are powerful to the extent that learning the facts that explain an explanandum would allow us to predict the occurrence the explanandum, if we didn't know that it had occurred. This assumption is made explicit in attempts to formalize explanatory power due to Schupbach and Sprenger (2011) and Crupi and Tentori (2012). Moreover, Eva and Stern (2019) provide a specific formalization of the explanatory power of *causal* explanations by assuming that, all else being equal, a causal explanation is powerful to the extent that learning that an *intervention* has brought about a particular cause of an event would allow us to predict the occurrence of that event. Let us call this feature of a causal explanation its "causal statistical relevance."

These two putative good-making features of a causal explanation can be in tension with one another. Consider the following example, due to Eva and Stern (2019):

> **Ettie:** 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 (p. 1047-8).

When Ettie's father explains the team's loss by their being down fifty points at the start of the fourth quarter, he provides an explanation with high causal statistical relevance; given an intervention on the game such that the local team is down fifty points at the end of the fourth quarter, it is very likely that they will lose. However, Ettie balks at the explanation because it has very low causal history depth; we don't get much of a story as to why the local team lost. Indeed, it is only once Ettie's father cites more distant causal factors contributing to the team's loss that Ettie's curiosity is satisfied.

The goal of this paper is to formalize the desiderata that an explanation is good to the extent that it possesses causal history depth and causal statistical relevance, and then prove a representation theorem showing that a specific family of functions provides a measure of causal explanatory power that uniquely satisfies both causal history depth and causal statistical relevance, alongside some minimal ancillary desiderata. My formalization uses the Bayesian network approach to causal representation and the interventional calculus found in the work of Pearl (2000). The result shows that we can mathematically represent a notion of causal explanatory power in which the overall quality of a causal explanation involves a trade-off between causal history depth and causal statistical relevance.

2. Formal Preliminaries

2.1. Bayesian Networks.

We begin with a probability space $\mathcal{P} = (\Omega, \Sigma, \Pr)$, where Ω is a sample space of primitive possibilities, Σ is an algebra on Ω (i.e., a set of subsets closed under union, complement, and intersection), and Pr is a probability distribution on Σ. A **random variable** $V : \Omega \rightarrow R_V$ is any function from the sample space into some range. For the purposes of this paper, I consider only random variables with finite ranges. A random variable V is said to be **measurable** with respect to a probability space $\mathcal{P} = (\Omega, \Sigma, \Pr)$ if and only if, for every $v \in R_V$, $V^{-1}(v) \in \Sigma$. This allows us to assign a probability to the variable taking any value in its range, using the formula $\Pr(V = v) = \Pr(V^{-1}(v))$.

Moving to the Bayesian network approach to the representation of the causal structure of a data-generating processs, I begin with the following definition:

Definition 2.1. A **causal graph** is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, where $\mathcal{V}$ is a set of random variables that are each measurable with respect to a common probability space $\mathcal{P} = (\Omega, \mathcal{A}, \Pr)$, and $\mathcal{R}$ is an acyclic set of ordered pairs of elements of $\mathcal{V}$, usually represented pictorially as arrows from one random variable to another.

The fundamental idea behind the Bayes nets approach to representing causal structure is that if there is a chain of arrows from one variable to another, then the first variable is causally relevant to the second. So, for instance, in an epidemiological causal graph there might be a chain of arrows from a variable representing whether or not a patient smokes to a variable representing whether or not the patient develops lung cancer, thus encoding the claim that smoking causes lung cancer.

If there is an arrow from one variable to another, we say that the first variable is **parent** of the second, and the second variable is a *child* of the first. We can then define the **ancestor** and **descendant** relations as the transitive closure of the parent and child relations, respectively. We are now in a position to define the all-important Markov condition:

Definition 2.2. A probability distribution Pr is **Markov** with respect to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, where all variables in $\mathcal{V}$ are measurable with respect to some probability space $\mathcal{P} = (\Omega, \mathcal{A}, \Pr)$, if and only if, according to Pr, each $\mathbf{X} \subseteq \mathcal{V}$ is independent of any subset of the set of non-descendants of $\mathbf{X}$ in $\mathcal{G}$, conditional on its parents in $\mathcal{G}$.

The Markov condition ensures that once we know the value taken by the direct causes of some variable set $\mathbf{X}$, information about the values taken by any non-effects of $\mathbf{X}$ are uninformative with respect to the probability that $\mathbf{X}$ takes any value. This reflects the intuitive condition that once we know the direct causes of $\mathbf{X}$, information about more distant causes of $\mathbf{X}$, or about other phenomena not causally related to $\mathbf{X}$, should not be relevant for making predictions about $\mathbf{X}$.

Finally, we are in a position to define a Bayesian network:

Definition 2.3. A **Bayesian network (or, "Bayes net")** is a pair $(\mathcal{G}, \Pr)$ such that $\mathcal{G}$ is a graph in which all variables in $\mathcal{V}$ are measurable with respect to some probability space $\mathcal{P} = (\Omega, \mathcal{A}, \Pr)$, no variable is an ancestor of itself (i.e., the graph is acyclic), and Pr is Markov to $\mathcal{G}$.

The core idea of the theory of causal Bayes nets is that, for the reasons given above, the causal structure of any system can be represented as a Bayes net $(\mathcal{G}, \Pr)$. To illustrate, consider the simple causal graph $X \rightarrow Y \rightarrow Z \leftarrow W$. If this graph can be paired with the probability distribution Pr in order to form a Bayes net, then it must be the case that, according to Pr, X is unconditionally independent of W, Y is independent of W conditional on X, Z is independent of X conditional on Y and W, and W is unconditionally independent of X and Y. These independence claims are individually necessary and jointly sufficient for Pr being Markov to the graph.

2.2. Intervention Distributions. Representing the causal structure of a system as a Bayes net allows us to calculate the probability distribution over a variable in that Bayes net, given an *intervention* on the system. An intervention is an exogenous setting of the values of one or more variables in the Bayes net that does not depend on values taken by any of the other variables. To see how this works, let us begin with a result from Pearl (2000, p. 15-16), who proves that if $\mathcal{V} = \{V_1, \ldots, V_m\}$ is the set of variables in a Bayes net and if each variable in $\mathcal{V}$ has a corresponding value $v_1, \ldots, v_m$, and if $\text{par}(V_i)$ is the vector of values taken by the set of parents of a variable V_i in the Bayes net, then the probability $\Pr(v_1, \ldots, v_m)$ can be factorized as follows.

$$\Pr(v_1, \ldots, v_m) = \prod_{i=1}^{m} \Pr(v_i | \text{par}(V_i)) \tag{2.1}$$

Next, suppose that we intervene on a set of variables $\mathbf{X} \subseteq \mathcal{V}$, setting it to the set of values $\mathbf{x}$. Pearl (2000, p. 30) and Spirtes et al. (2000, p. 51) show that in a Bayes net, the interventional conditional probability $\Pr(v_1, \ldots, v_m | do(\mathbf{x}))$ can be obtained using the following truncated factorization:

$$\Pr(v_1, \ldots, v_m | do(\mathbf{x})) = \prod_{i=1}^{m} \Pr_{do(\mathbf{x})}(v_i | \text{par}(V_i)) \tag{2.2}$$

Where each probability $\Pr_{do(\mathbf{x})}(v_i | \text{par}(V_i))$ is defined as follows:

$$\Pr_{do(\mathbf{x})}(v_i | \text{par}(V_i)) = \begin{cases} \Pr(v_i | \text{par}(V_i)) & \text{if } V_i \notin \mathbf{X} \\ 1 & \text{if } V_i \in \mathbf{X} \text{ and } v_i \text{ consistent with } \mathbf{x} \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Huang and Valtorta (2006) show that this procedure can be used to calculate the probability distribution over any combination of variable values in a Bayes net, given any intervention.

A less formal account of the connection between Bayes nets and interventional conditional probability distributions can be stated as follows. Let $(\mathcal{G}, \Pr)$ be a Bayes net. If we intervene on some set of variables $\mathbf{X} \subseteq \mathcal{V}$, then we make it the case that the values of the variables in $\mathbf{X}$ no longer depend on their parents, but instead depend solely on the intervention. This can be represented graphically by a sub-graph of $(\mathcal{G}, \Pr)$ in which all arrows into all variables in $\mathbf{X}$ are removed. This sub-graph is called the **pruned sub-graph** of $(\mathcal{G}, \Pr)$ for an intervention on $\mathbf{X}$. Spirtes, Glymour, and Scheines (2000) prove that $\Pr_{do(\mathbf{x})}$ will be Markov to this pruned sub-graph of $(\mathcal{G}, \Pr)$, so that we can calculate the joint probability distribution over the pruned sub-graph created by any intervention on any set of variables $\mathbf{X}$, using Eq. 2.2. This calculation allows us to determine which types of events represented in a Bayes net cause other types of events represented in the same Bayes net, since we can use interventions both to change the values of variables and to hold selected variables fixed at their actual values.

2.3. Causal Distance. In the introduction, I introduced the desideratum that, all else being equal, the more that a causal explanation cites the full causal history of an explanandum, tracing that history further back in a causal chain, the more powerful that causal explanation is. So, we will need to define a measure of causal history depth, in the context of a given Bayes net. Let us begin with some graph-theoretic terminology.

Definition 2.4. For any two variables X and Y in a graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, a **directed path** from X to Y is a set of edges $\{R_1, \ldots, R_n\}$ such that:

i. Each R_i in the set is an element of $\mathcal{R}$,
ii. $R_1 = (X, V_j)$, where $V_j \in \mathcal{V}$,
iii. $R_n = (V_k, Y)$, where $V_k \in \mathcal{V}$, and
iv. there exists a sequence of distinct variables $(V_1, \ldots, V_{n+1})$ such that for each R_i in the path, $R_i = (V_i, V_{i+1})$.

Pictorially, there is a directed path from X to Y in a graph if one can follow the edges of the graph to "travel" from X to Y, moving with the direction of the edges, without passing through the same variable more than once. To illustrate, in the graph $X \to Y \to Z \leftarrow W$, there is a directed path from X to Z, but not from X to W.

Using the cardinalities of directed paths between variables, we define a proximity measure measure on the variables in a graph, in two steps.

Definition 2.5. For any causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, the **causal distance** $\delta_{\mathcal{G}}(X, Y)$ between two variables $X \in \mathcal{V}$ and $Y \in \mathcal{V}$ is the cardinality of the directed path from X to Y with minimal cardinality, if such a directed path exists. If no such path exists, then $\delta_{\mathcal{G}}(X, Y) = 0$.

Definition 2.6. For any causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, the **normalized causal proximity** $\pi_{\mathcal{G}}(\mathbf{X}, \mathbf{Y})$ takes as its arguments any two sets $\mathbf{X} \subseteq \mathcal{V}$ and $\mathbf{Y} \subseteq \mathcal{V}$, and is defined as follows:

$$\pi_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}) = \frac{\max\{\delta_{\mathcal{G}}(V_i, V_j) : V_i, V_j \in \mathcal{V}\} - \max\{\delta_{\mathcal{G}}(X, Y) : X \in \mathbf{X}, Y \in \mathbf{Y}\}}{\max\{\delta_{\mathcal{G}}(V_i, V_j) : V_i, V_j \in \mathcal{V}\}}$$

In other words, $\pi_{\mathcal{G}}(\mathbf{X}, \mathbf{Y})$ returns the normalized difference between the length of the longest shortest directed path between any two variables in the graph $\mathcal{G}$ and the length of the longest shortest path between a variable in $\mathbf{X}$ and a variable in $\mathbf{Y}$. The result is a measure of proximity that approaches one as the longest shortest path between a variable in $\mathbf{X}$ and a variable in $\mathbf{Y}$ gets shorter in length, and approaches zero as the longest shortest path between a variable in $\mathbf{X}$ and a variable in $\mathbf{Y}$ becomes longer. To illustrate, in the graph $X \to Y \to Z \leftarrow W$, $\pi_{\mathcal{G}}(\{X\}, \{W\}) = 0$, $\pi_{\mathcal{G}}(\{Y, W\}, \{Z\}) = .5$, and $\pi_{\mathcal{G}}(\{X, Y\}, \{Z\}) = 0$. As it will occasionally be more convenient to speak in terms of normalized causal distance rather than normalized causal proximity, we define a normalized causal distance function $\Delta_{\mathcal{G}}(\mathbf{X}, \mathbf{Y})$:

Definition 2.7. For any causal graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$, the **normalized causal distance** $\Delta_{\mathcal{G}}(\mathbf{X}, \mathbf{Y})$ is given by the equation $\Delta_{\mathcal{G}}(\mathbf{X}, \mathbf{Y}) = 1 - \pi_{\mathcal{G}}(\mathbf{X}, \mathbf{Y})$.

3. The Representation Theorem

In this primary section of the paper, I make good on my promise in the introduction to state a set of desiderata that formalize those cases in which explanatory power requires a trade-off between causal history depth and predictive power, and then prove that a specific family of measures uniquely satisfies these desiderata. In several respects, my proposed desiderata are adapted from those proposed by Schupbach and Sprenger (2011), but with adaptations made so as to incorporate causal history depth and intervention distributions, neither of which Schupbach and Sprenger consider.

I begin by stating three ancillary desiderata for such a measure. The first is as follows:

> **D1 (Formal Structure)**. For any Bayes net $(\mathcal{G}, \Pr)$ where the graph $\mathcal{G} = (\mathcal{V}, \mathcal{R})$ is such that each variable in $\mathcal{V}$ is measurable with respect to the probability space $\mathcal{P} = (\Omega, \Sigma, \Pr)$, $\theta_{\mathcal{P},\mathcal{G}}$ is a function from any two sets of values $\mathbf{e}$ and $\mathbf{c}$ of any two sets of variables $\mathbf{E} \subseteq \mathcal{V}$ and $\mathbf{C} \subseteq \mathcal{V}$ to a real number $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c}) \in [-1, 1]$ that can be represented as a function of $\Pr(\mathbf{e}|do(\mathbf{c}))$, $\Pr(\mathbf{e})$, and $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$.

This desideratum ensures that $\theta_{\mathcal{P},\mathcal{G}}$ takes as input: i) the fact that a set of effect variables $\mathbf{E}$ takes a set of values $\mathbf{e}$, and ii) the fact that a set of causal variables $\mathbf{C}$ takes a set of values $\mathbf{c}$, and returns a value between -1 and 1 representing the power with which the fact that $\mathbf{C} = \mathbf{c}$ explains the fact that $\mathbf{E} = \mathbf{e}$. Moreover, this value is determined solely by the following quantities: i) the probability that $\mathbf{E} = \mathbf{e}$ given an intervention setting $\mathbf{C}$ to $\mathbf{c}$, ii) the marginal probability that $\mathbf{E} = \mathbf{e}$, and iii) the normalized causal proximity between $\mathbf{C}$ and $\mathbf{E}$.

Second, I introduce an additional formal constraint:

> **D2 (Normality and Form).** The function $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c})$ is a ratio of two functions of $\Pr(\mathbf{e}|do(\mathbf{c}))$, $\Pr(\mathbf{e})$, and $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$, each of which are homogeneous in their arguments to lowest possible degree $k \geq 1$.

The requirement that the function be a ratio of two functions with the same arguments ensures that it is normalized. I follow Schupbach and Sprenger in holding that requiring that each function be homogenous in its arguments to lowest possible degree $k \geq 1$ ensures that their measure of explanatory power is maximally *simple*, in a well-defined sense advocated by Carnap (1950) and Kemeny and Oppenheim (1952). Note that a function f is homogeneous in its arguments $x_1, \ldots, x_n$ to degree k if for all $\gamma \in \mathbb{R}$, $f(\gamma x_1, \ldots, \gamma x_n) = \gamma^k f(x_1, \ldots, x_n)$.

Third, I introduce a desideratum aimed at capturing the idea that there is a specific zero point for any measure of explanatory power:

> **D3 (Neutrality).** If $\Pr(\mathbf{e}|do(\mathbf{c})) = \Pr(\mathbf{e})$, then $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c}) = 0$.

Neutrality ensures that when an intervention setting causal variables to a particular set of values provides no information about the explanandum effect, causal explanatory power is zero.

With these three ancillary desiderata established, I move now to a formalization of causal history depth:

> **D4 (Causal History Depth).** Holding fixed the value of $\Pr(\mathbf{e}|do(\mathbf{c}))$ and $\Pr(\mathbf{e})$, if $\Pr(\mathbf{e}|do(\mathbf{c})) > \Pr(\mathbf{e})$, then $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c})$ is strictly decreasing in $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$, and if $\Pr(\mathbf{e}|do(\mathbf{c})) < \Pr(\mathbf{e})$, then $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c})$ is strictly increasing in $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$.

This desideratum encodes the idea that, all else being equal, if an intervention setting $\mathbf{C}$ to $\mathbf{c}$ is positively statistically relevant to the event denoted by $\mathbf{E} = \mathbf{e}$, then $\mathbf{C} = \mathbf{c}$ is explanatorily powerful to the extent that it cites causes that are more causally distant (and so, less proximal) with respect to the variables in $\mathbf{E}$. Moreover, it introduces the idea that if an intervention setting $\mathbf{C}$ to $\mathbf{c}$ is *negatively* statistically relevant to the event denoted by $\mathbf{E} = \mathbf{e}$, then explanatory power is an increasing function of causal history depth (and so, a decreasing function of causal proximity). This reflects the assumption that attempted explanations that cite factors that both make the event being explained less likely and are causally far removed from the event being explained are especially bad explanations.

Fifth and finally, I introduce a formalization of causal statistical relevance:

> **D5 (Causal Statistical Relevance).** Holding fixed the value of $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$, the greater the degree of causal statistical relevance between $\mathbf{e}$ and $\mathbf{c}$ (defined here as the difference $\Pr(\mathbf{e}|do(\mathbf{c})) - \Pr(\mathbf{e})$), the greater the value of $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c})$.

This desideratum says that the more an intervention such that $\mathbf{C} = \mathbf{c}$ makes it likely that $\mathbf{E} = \mathbf{e}$, the greater the explanatory power of $\mathbf{c}$ with respect to $\mathbf{e}$.

These five desiderata together determine the form of a more general measure of causal explanatory power, as established by the following representation theorem (see appendix for a proof of this and all subsequent facts and propositions):

Proposition 3.1. *Any measure $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c})$ that satisfies **D1-D5** has the form:*

$$\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c}) = \frac{\Pr(\mathbf{e}|do(\mathbf{c})) - \Pr(\mathbf{e})}{\Pr(\mathbf{e}|do(\mathbf{c})) + \Pr(\mathbf{e}) + \alpha\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})} \textit{ where } \alpha > 0.$$

The equation for $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c})$ can be re-written in terms of normalized causal distance as follows:

$$\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c}) = \frac{\Pr(\mathbf{e}|do(\mathbf{c})) - \Pr(\mathbf{e})}{\Pr(\mathbf{e}|do(\mathbf{c})) + \Pr(\mathbf{e}) + \alpha[1 - \Delta_{\mathcal{G}}(\mathbf{C}, \mathbf{E})]} \text{ where } \alpha > 0. \tag{3.1}$$

This result raises the immediate question of the significance of the coefficient α. For a given Bayes net $(\mathcal{G}, \Pr)$ with variable settings $\mathbf{C} = \mathbf{c}$ and $\mathbf{E} = \mathbf{e}$, let $\phi_{\mathcal{P},\mathcal{G}}$ be a function defined as follows:

$$\phi_{\mathcal{P},\mathcal{G}}(\alpha; \mathbf{e}, \mathbf{c}) = \frac{\left|\frac{\partial\theta_{\mathcal{P},\mathcal{G}}(\mathbf{c},\mathbf{e})}{\partial\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E})}\right|}{\left|\frac{\partial\theta_{\mathcal{P},\mathcal{G}}(\mathbf{c},\mathbf{e})}{\partial\Pr(\mathbf{e}|do(\mathbf{c}))}\right|}. \tag{3.2}$$

If we take the absolute value of the partial derivative of $\theta_{\mathcal{P},\mathcal{G}}$ with respect to any argument to measure the importance of that argument to the overall measure of causal explanatory power, then $\phi_{\mathcal{P},\mathcal{G}}$ measures the relative importance of causal proximity/distance, as compared to the statistical relevance of an intervention setting $\mathbf{C}$ to $\mathbf{c}$, for a fixed value of $\Pr(\mathbf{e})$.[1] The following fact about $\phi_{\mathcal{P},\mathcal{G}}$ holds:

Fact 3.2. For any Bayes net $(\mathcal{G}, \Pr)$ and any $\Pr(\mathbf{e}|do(\mathbf{c}))$, $\Pr(\mathbf{e})$, and $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$, if $\Pr(\mathbf{e}|do(\mathbf{c})) \neq \Pr(\mathbf{e})$, then $\frac{d\phi_{\mathcal{P},\mathcal{G}}(\alpha;\mathbf{e},\mathbf{c})}{d\alpha} > 0$.

[1] There is a slight idealization at work here. In practice, $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$ can only take rational values in the unit interval, and so the partial derivative $\frac{\partial\theta_{\mathcal{P},\mathcal{G}}}{\partial\theta_{\mathcal{G}}(\mathbf{C},\mathbf{E})}$ is not really well-defined. However, for the purpose of calculating $\phi_{\mathcal{P},\mathcal{G}}$, we treat $\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$ as though it can take all real values in the unit interval.

Thus, increases in α result in increases in the relative importance of proximity/distance, as compared to causal statistical relevance, for the measure of causal explanatory power, whenever there is some causal statistical relevance, either positive or negative, between the explanans and the explanandum.

To illustrate how this measure works, let us return to Eva and Stern's **Ettie** example:

Example 3.3. Consider the simple causal graph $X \rightarrow Y \rightarrow Z$, where X is a binary variable denoting whether or not the team's best player is injured in the first half of the match (0 if not injured, 1 if injured), Y is a binary variable denoting whether or not the home team is down by more than thirty points at the start of the fourth quarter (0 if they are not, 1 if they are), and Z is a binary variable denoting whether or not the home team loses (0 if they lose, 1 if they do not lose). Suppose that $\Pr(Z=0|do(X=1)) = .8$, $\Pr(Z=0|do(Y=1)) = .99$, and $\Pr(Z=0) = .3$. We know that $\Delta_{\mathcal{G}}(\{X\},\{Z\}) = 1$ and $\Delta_{\mathcal{G}}(\{Y\},\{Z\}) = .5$. It follows that if $\alpha > .456$, then $\theta_{\mathcal{P},\mathcal{G}}(Z=0, X=1) > \theta_{\mathcal{P},\mathcal{G}}(Z=0, Y=1)$.

Thus, for suitably large α (and so, suitably large emphasis on causal history depth as a determinant of causal explanatory power), my proposed measure of causal explanatory power can deliver verdicts in keeping with Ettie's intuitions in this vignette.

One might object at this stage that the formalization of causal history depth presented here only tracks the degree to which an explanation cites a distant cause relative to the explanandum effect, and that this is distinct from the desideratum that an explanation fills in the full causal history of the events leading up to the explanandum effect. In response, I prove a result showing that, necessarily, the function derived above will deliver the result that the explanatory power of a causal explanation is always positively associated with the extent to which that explanation cites the full causal history of an explanandum effect.

Consider any Bayes net $\mathcal{G} = (\mathcal{V}, \mathcal{R})$ in which all variables are measurable with respect to some probability space $\mathcal{P}$. Let $\mathbf{E}$ be some subset of $\mathcal{V}$, and let $\mathrm{Par}_0(\mathbf{E})$ denote the parents of the variables in $\mathbf{E}$ according to $\mathcal{G}$, let $\mathrm{Par}_1(\mathbf{E})$ denote the parents of the parents of the variables in $\mathbf{E}$ according to $\mathcal{G}$, and so on. Let $\Xi(n) = \bigcup_{i=0}^{n} \mathrm{Par}_i(\mathbf{E})$, and let $\xi(n)$ be a set of values taken by the variables in $\Xi(n)$. The following proposition holds:

Proposition 3.4. *For all $n > 0$, if $\mathrm{Par}_n(\mathbf{E})$ is non-empty, $\mathrm{Par}_n(\mathbf{E}) \neq \mathrm{Par}_{n-1}(\mathbf{E})$, and $\mathbf{E} \cap \Xi(n) = \emptyset$, then $\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \xi(n)) > \theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \xi(n-1))$.*

This ensures that, for any set of variables $\mathbf{E}$, we can generate a more powerful explanation of why $\mathbf{E}$ takes the value that it does by accounting for more of the causal history of the event represented by $\mathbf{E} = \mathbf{e}$. This shows that when we stipulate as desiderata for a measure of causal explanatory power my formalizations of causal history depth and causal statistical relevance, the measure captures the idea that, all else being equal, ideal causal explanation involves a maximally perspicuous filling-in of the causal chain of events resulting in the explanandum effect, in keeping with the motivating intuition of this paper.

4. Conclusion

I conclude by first noting that my goal in this paper has *not* been to give a formal measure of causal explanatory power that delivers intuitive judgements in all applicable circumstances. Indeed, I take it that no all-things-considered quantitative measure of explanatory power could possibly comport with our intuitions or scientific practices in all cases.[2] Instead, my aim has been to examine specifically those cases in which the power of a causal explanation is determined by a trade-off between causal history depth and causal statistical relevance.

Even with this qualification, it could be argued that there is no context in which explanatory power is *entirely* determined by a trade-off between these two properties, and that instead there is always a wide array of factors that determine causal explanatory power in any given context, such that the concept of explanatory power itself never admits of formal representation. Against this line of argument, I hold that in some cases, the sole *primary* determinants of causal explanatory power are causal history depth and causal statistical relevance. In these cases, my measure amounts to an *explication* of explanatory power, in the sense of Carnap (1950). That is, it takes an inherently vague, imprecise notion from the real world and renders it mathematically tractable, while still capturing something close enough to the actual determinants of our judgements of explanatory power.

A further avenue for future work would be an empirical investigation of the conditions under which human beings actually trade-off causal distance against statistical relevance when assessing the explanatory virtues of causal explanations. Having found these conditions, we could then investigate the dynamics of these trade-offs, and the extent to which my proposed measure actually predicts human judgments. From a theoretical point of view, there is also more work to be done examining the various properties and implications of the measure of causal explanatory power proposed here, and to provide a more in-depth comparison with alternative measures of causal explanatory power, most notably that provided in Eva and Stern (2019).

[2] See Lange (forthcoming) for an argument to this effect.

References

Carnap, Rudolf. *Logical foundations of probability.* University of Chicago Press, 1950.

Crupi, Vincenzo and Katya Tentori. "A second look at the logic of explanatory power (with two novel representation theorems)". In: *Philosophy of Science* 79.3 (2012), pp. 365–385.

Eva, Benjamin and Reuben Stern. "Causal explanatory power". In: *The British Journal for the Philosophy of Science* 70.4 (2019), pp. 1029–1050.

Huang, Yimin and Marco Valtorta. "Pearl's calculus of intervention is complete". In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence.* 2006, pp. 217–224.

Keas, Michael N. "Systematizing the theoretical virtues". In: *Synthese* 195.6 (2018), pp. 2761–2793.

Kemeny, John G and Paul Oppenheim. "Degree of factual support". In: *Philosophy of Science* 19.4 (1952), pp. 307–324.

Lange, Marc. "Against Probabilistic Measures of Explanatory Quality". In: *Philosophy of Science* (forthcoming). URL: http://philsci-archive.pitt.edu/20205/.

Lewis, David. "Causal Explanation". In: *Philosophical Papers Vol. II.* Ed. by David Lewis. Oxford University Press, 1986, pp. 214–240.

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

Railton, Peter. "Probability, explanation, and information". In: *Synthese* (1981), pp. 233–256.

Salmon, Wesley C. *Scientific Explanation and the Causal Structure of the World.* Princeton University Press, 1984.

Schupbach, Jonah N. and Jan Sprenger. "The Logic of Explanatory Power". In: *Philosophy of Science* 78.1 (2011), pp. 105–127.

Spirtes, Peter, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* Mit Press: Cambridge, 2000.

Appendix A. Proofs and Demonstrations

A.1. Proof of Prop. 3.1.

Proof. For the sake of concision, let $x = \Pr(\mathbf{e}|do(\mathbf{c}))$, let $y = \Pr(\mathbf{e})$, and let $z = \pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})$. By **D1**, a measure of causal explanatory power must be a function $f(x, y, z)$. We begin by searching for a function that is homogenous in its arguments to degree 1, in keeping with **D2**. Such a function has the form

$$f(x, y, z) = \frac{ax + by + cz}{\bar{a}x + \bar{b}y + \bar{c}z}. \tag{A.1}$$

D3 requires that the numerator is zero whenever $x = y$. This is achieved by letting $a = -b$ and $c = 0$, so that we have:

$$f(x, y, z) = \frac{a(x - y)}{\bar{a}x + \bar{b}y + \bar{c}z}. \tag{A.2}$$

Letting $x = 1$ gives us:

$$f(x, y, z) = \frac{a - ay}{\bar{a} + \bar{b}y + \bar{c}z}. \tag{A.3}$$

By **D1**, **D4** and **D5**, as $y \to 0$ and $z \to 0$, it must be the case that $f(x, y, z) \to 1$. This requires that $a = \bar{a}$, so that we have:

$$f(x, y, z) = \frac{a(x - y)}{ax + \bar{b}y + \bar{c}z}. \tag{A.4}$$

Next, let $x = 0$ so that we have:

$$f(x, y, z) = \frac{-ay}{\bar{b}y + \bar{c}z}. \tag{A.5}$$

By **D1**, **D4** and **D5**, as $y \to 1$ and $z \to 0$, it must be the case that $f(x, y, z) \to -1$. This requires that $\bar{b} = a$, so that we have:

$$f(x, y, z) = \frac{a(x - y)}{a(x + y) + \bar{c}z}. \tag{A.6}$$

It remains to determine the sign of a and $\bar{c}$. Let $x = 1$ and $y = 0$, so that

$$f(x, y, z) = \frac{a}{a + \bar{c}z}. \tag{A.7}$$

If $\bar{c} < 0$, then $f(x, y, z) > 1$ for positive z, in violation of **D1**. Thus, $\bar{c} \geq 0$. Moreover, it must be the case that $\bar{c} > 0$ for **D4** to hold in general. Next, let $x = 0$ and $y = 1$, so that:

$$f(x, y, z) = \frac{-a}{a + \bar{c}z}. \tag{A.8}$$

If $a < 0$, then $f(x, y, z) < -1$ for all $\bar{c} > 0$, in violation of **D1**. Thus, $a \geq 0$. Moreover, it must be the case that $a > 0$ for **D5** to hold in general. Letting $\alpha = \frac{\bar{c}}{a}$ we arrive at the function:

$$f(x, y, z) = \frac{x - y}{x + y + \alpha z}, \tag{A.9}$$

or:

$$\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e}, \mathbf{c}) = \frac{\Pr(\mathbf{e}|do(\mathbf{c})) - \Pr(\mathbf{e})}{\Pr(\mathbf{e}|do(\mathbf{c})) + \Pr(\mathbf{e}) + \alpha\pi_{\mathcal{G}}(\mathbf{C}, \mathbf{E})} \text{ where } \alpha > 0. \tag{A.10}$$

□

A.2. Demonstration of Fact 3.2.

Proof. We proceed by expanding the function $\phi_{\mathcal{P},\mathcal{G}}$:

$$\phi_{\mathcal{P},\mathcal{G}}(\alpha;\mathbf{e},\mathbf{c}) = \frac{\left|\frac{\partial\theta_{\mathcal{P},\mathcal{G}}(\mathbf{e},\mathbf{c})}{\partial\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E})}\right|}{\left|\frac{\partial\theta_{\mathcal{P},\mathcal{G}}}{\partial\Pr(\mathbf{e}|do(\mathbf{c}))}\right|} \tag{A.11}$$

$$\phi_{\mathcal{P},\mathcal{G}}(\alpha;\mathbf{e},\mathbf{c}) = \frac{\left|\frac{-\alpha[\Pr(\mathbf{e}|do(\mathbf{c}))-\Pr(\mathbf{e})]}{(\Pr(\mathbf{e}|do(\mathbf{c}))+\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E}))^2}\right|}{\left|\frac{2\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E})}{(\Pr(\mathbf{e}|do(\mathbf{c}))+\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E}))^2}\right|} \tag{A.12}$$

Since all terms are positive, if $\Pr(\mathbf{e}|do(\mathbf{c})) > \Pr(\mathbf{e})$, then we have

$$\phi_{\mathcal{P},\mathcal{G}}(\alpha;\mathbf{e},\mathbf{c}) = \frac{\alpha[\Pr(\mathbf{e}|do(\mathbf{c}))-\Pr(\mathbf{e})]}{2\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E})}, \tag{A.13}$$

in which case

$$\frac{d\phi_{\mathcal{P},\mathcal{G}}(\alpha;\mathbf{e},\mathbf{c})}{d\alpha} = \frac{2\Pr(\mathbf{e})[\Pr(\mathbf{e}|do(\mathbf{c}))-\Pr(\mathbf{e})]}{(2\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E}))^2} > 0 \tag{A.14}$$

If $\Pr(\mathbf{e}|do(\mathbf{c})) < \Pr(\mathbf{e})$, then we have

$$\phi_{\mathcal{P},\mathcal{G}}(\alpha;\mathbf{e},\mathbf{c}) = \frac{-\alpha[\Pr(\mathbf{e}|do(\mathbf{c}))-\Pr(\mathbf{e})]}{2\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E})}, \tag{A.15}$$

in which case

$$\frac{d\phi_{\mathcal{P},\mathcal{G}}(\alpha;\mathbf{e},\mathbf{c})}{d\alpha} = \frac{-2\Pr(\mathbf{e})[\Pr(\mathbf{e}|do(\mathbf{c}))-\Pr(\mathbf{e})]}{(2\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\mathbf{C},\mathbf{E}))^2} > 0. \tag{A.16}$$

Thus, the fact holds in either case. □

A.3. Proof of Prop. 3.4.

Proof. Since $\mathrm{Par}_n(\mathbf{E}) \neq \mathrm{Par}_{n-1}(\mathbf{E})$, we know that at there is at least one $X \in \mathrm{Par}_n(\mathbf{E})$ such that $\delta_{\mathcal{G}}(X,E) > \delta_{\mathcal{G}}(Y,E)$ for any $E \in \mathbf{E}$ and any $Y \in \mathrm{Par}_{n-1}(\mathbf{E})$. This entails that

$$\max\{\delta_{\mathcal{G}}(X,E) : X \in \mathrm{Par}_n(\mathbf{E}), E \in \mathbf{E}\} > \max\{\delta_{\mathcal{G}}(X,E) : X \in \mathrm{Par}_{n-1}(\mathbf{E}), E \in E\},$$

which entails in turn that

$$\max\{\delta_{\mathcal{G}}(X,E) : X \in \Xi(n), E \in \mathbf{E}\} > \max\{\delta_{\mathcal{G}}(X,E) : X \in \Xi(n-1), E \in E\},$$

and so $\pi_{\mathcal{G}}(\Xi(n),\mathbf{E}) < \pi_{\mathcal{G}}(\Xi(n-1),\mathbf{E})$. Since $\mathbf{E}$ and $\Xi(n)$ have empty intersection, we know from Eq. 2.3 that for any $\mathbf{e}$, $\Pr(\mathbf{e}|do(\xi(n))) = \Pr(\mathbf{e}|\operatorname{par}_0(\mathbf{E}))$ for any n, and so, for $n > 0$, $\Pr(\mathbf{e}|do(\xi(n))) = \Pr(\mathbf{e}|do(\xi(n-1))) = \Pr(\mathbf{e}|\operatorname{par}_0(\mathbf{E}))$. Together, this entails that, for any α,

$$\begin{aligned}\theta_{\mathcal{P},\mathcal{G}}(\mathbf{E},\Xi(n)) &= \frac{\Pr(\mathbf{e}|do(\xi(n)))-\Pr(\mathbf{e})}{\Pr(\mathbf{e}|do(\xi(n)))+\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\Xi(n),\mathbf{E})}\\ &> \frac{\Pr(\mathbf{e}|do(\xi(n-1)))-\Pr(\mathbf{e})}{\Pr(\mathbf{e}|do(\xi(n-1)))+\Pr(\mathbf{e})+\alpha\pi_{\mathcal{G}}(\Xi(n-1),\mathbf{E})} = \theta_{\mathcal{P},\mathcal{G}}(\mathbf{E},\Xi(n-1)).\end{aligned}$$

□