# Toward a Formal Pragmatics of Explanation

**Jacqueline Harding** (hardingj@stanford.edu), Department of Philosophy, Stanford University

**Tobias Gerstenberg** (gerstenberg@stanford.edu), Department of Psychology, Stanford University

**Thomas Icard** (icard@stanford.edu), Department of Philosophy, Stanford University

## Abstract

This paper presents a formal account of causal explanation, grounded in a theory of conversational pragmatics, and inspired by the interventionist idea that explanation is about asking and answering what-if-things-had-been-different questions. We illustrate the fruitfulness of the account, relative to previous accounts, by showing that widely recognized "explanatory virtues" emerge naturally, as do subtle empirical patterns concerning the impact of norms on causal judgments. An extended version of the paper with further details can be found here: https://arxiv.org/pdf/2505.03732.

**Keywords:** explanation, causation, pragmatics, speech acts

## Introduction

It is tempting to think of the philosophy of explanation and the psychology of explanation as separate subjects. Philosophers elevate good explanation as the epitome of good science. In contrast to "mere description," explanation helps reveal fundamental structures underlying observed empirical phenomena. Psychologists, meanwhile, focus on more "intuitive" modes of thought, highlighting the important role of explanatory cognition in the ways people spontaneously learn about the world and teach each other about it. While the philosophical literature has exerted considerable influence on the way psychologists investigate explanation (Lombrozo, 2006; Keil, 2006; Goddu & Gopnik, 2024), philosophers have been more reluctant to incorporate insights from psychology and cognitive science into their accounts.

As often highlighted, explanations can be conceived as answers to "why?" questions (Bromberger, 1965). For instance, asking "Why FACT?" (where FACT is some explanandum) is typically understood relative to a set of relevant alternatives to FACT. This set of alternatives will often depend on contextual factors, and such factors may affect both what is being asked, and what would be considered a good answer (Van Fraassen, 1980; Woodward, 2003; Ylikoski, 2007). An important strand of exploration of "pragmatic" theories of explanation has stressed the relevance of conversational and cognitive considerations (Bromberger, 1965; Van Fraassen, 1977, 1980; Achinstein, 1983; Ylikoski & Kuorikoski, 2010; Potochnik, 2017; De Regt, 2017). Yet, these proposals have mostly remained programmatic and informal.

Meanwhile, proponents of more formal philosophical accounts have not been unaware of these pragmatic and contextual factors, but have regarded them as peripheral to what is most interesting about explanation. The focus is instead on carving some distinctive "explanatory relation," analyzed in terms of sentences closed under logical deduction (Hempel & Oppenheim, 1948), statistical models and probability distributions (Douven, 2022; Hempel, 1965; Salmon, 1971), or physical models of spacetime (Dowe, 1992; Salmon, 1998).

One formal account that does incorporate some pragmatic factors is due to Gärdenfors (1980, 1988, 1990). A fundamental move in this line of work is to put the *consumer* of the explanation, and his *epistemic state*, front and center. It is assumed that facts demanding explanation are those which are surprising to the individual, in the sense that, prior to learning, he assigned them relatively low probability. For Gärdenfors, relativity to a listener's knowledge state is central.

A theme that runs through much of the literature on pragmatic approaches is that there is nothing fundamentally different about explanation above and beyond "mere" description: there is no distinctively *explanatory* relation. What is special about explanation is the way in which appropriate descriptive information resolves an agent's uncertainty through interaction with pragmatic context.

## Explanation and Causation

While conversational context is no doubt important, a common view is that explanation fundamentally involves relations of asymmetric dependence, chiefly causal dependence. The basic thought is that, e.g., telling someone that you took a walk in the park today (a description of *what* happened) is not typically an explanation, unless it helps address a causal-explanatory question (e.g., *why* your shoes were muddy).

Recent decades have seen a variety of causal frameworks centred around the notion of *causal intervention* (Spirtes, Glymour, & Scheines, 1993; Pearl, 1995; Woodward, 2003; Hitchcock, 2001; Sloman & Lagnado, 2015; Goddu & Gopnik, 2024; Rottman & Hastie, 2016; Waldmann, 2017; Lombrozo, 2006). Roughly speaking, an intervention on a system is an idealized manipulation of some component of the system, which leaves all other components unchanged. This has been made more precise with the help of mathematical models, among the most general of which is the *structural causal model* (Pearl, 2009; Peters, Janzing, & Schölkopf, 2017; Bareinboim, Correa, Ibeling, & Icard, 2022):

**Definition 1.** A structural causal model (SCM) $\mathcal{M}$ is a 4-



In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society* 

tuple $(\mathbf{U}, \mathbf{V}, \mathbf{f_V}, P(\mathbf{U}))$, where $\mathbf{U}$ is a set of *exogenous variables*, with possible values $\mathsf{Val}(U)$ for each $U \in \mathbf{U}$; $\mathbf{V}$ is a set of *endogenous variables*, with possible values $\mathsf{Val}(X)$ for each $X \in \mathbf{V}$; $\mathbf{f_V}$ is a set of *structural functions*, where $f_X : \mathsf{Val}(\mathbf{V} \cup \mathbf{U}) \rightarrow \mathsf{Val}(X)$ for each endogenous variable $X \in \mathbf{V}$; $P(\mathbf{U})$ is a probability distribution over $\mathsf{Val}(\mathbf{U})$.

The notion of *intervention* on an SCM is captured by mechanism replacement: intervening to set variable $X$ to value $x$ is a matter of replacing the structural function $f_X$ with the constant function sending all arguments to $x$.

How exactly do causal interventions fit into a broader theory of explanation? Most proponents of interventionist approaches to explanation agree that it has something to do with answering "what-if-things-had-been-different questions" (Woodward & Hitchcock, 2003). While this one move already marks a significant departure from previous approaches, it also leaves open many important details. Which what-if-things-had-been-different questions are relevant? And how are these questions to be answered?

**Halpern and Pearl on Explanation** Perhaps the most influential formalization of the interventionist approach is from Halpern and Pearl (2005b). They present a variant of Gärdenfors's epistemic analysis, with one new ingredient: appeal to a notion of *actual causation* (Halpern & Pearl, 2005a). Suppose some token event $E$ occurs. An "actual cause" of $E$ is a factor $C$ that causally contributed to $E$'s occurrence. A common way to think about causal contribution is counterfactual, employing the "but-for" test: if $E$ would not have occurred *but for* $C$, we say that $C$ was an actual cause of $E$.

Sometimes the but-for test fails. For instance, if two individually sufficient events both contributed to $E$ (if they "overdetermined" $E$), then $E$ still would have occurred even if one of the causes had been absent. This motivates more sophisticated counterfactual analyses. The rough idea is that $C$ should be a but-for cause under some contingency in which we hold the values of some variables (off a "causal pathway" from $C$ to $E$) fixed to some (possibly non-actual) values. Different proposals make this intuition precise in subtly different ways (Hitchcock, 2001; Woodward, 2003; Halpern & Pearl, 2005a; Gallow, 2021). Such nuances will not matter for our purposes; we only assume that we have fixed some "egalitarian" analysis of what it is to be a causally contributing factor.

Concretely, Halpern and Pearl (2005b) suppose that explanation is relative to a person's epistemic state $\mathcal{K}$, a set of pairs $(\mathcal{M}, \mathbf{u})$, where $\mathcal{M}$ is a causal model and $\mathbf{u}$ is a "context," i.e., values for all the exogenous variables. Intuitively, these are the causal situations that are consistent with the person's knowledge. It is supposed for simplicity that all models in $\mathcal{K}$ have the same variables, so that uncertainty is only over the structural relationships and the values of (both exogenous and endogenous) variables. Halpern and Pearl (2005b) propose:

**Definition 2** (HP Analysis of Explanation)**.** $\mathbf{X} = \mathbf{x}$ is an explanation of $\mathsf{FACT}$ relative to an epistemic state $\mathcal{K}$ iff the following conditions hold:

**(EX1)** $\mathsf{FACT}$ is true at all models $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$.

**(EX2)** $\mathbf{X} = \mathbf{x}$ is an actual cause of $\mathsf{FACT}$ for all $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$ in which $\mathbf{X} = \mathbf{x}$ is true.

**(EX3)** No proper subset of $\mathbf{X}$ satisfies **EX2**.

**(EX4)** There exists $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$ in which $\mathbf{X} \neq \mathbf{x}$.

The agent is certain that some explanandum holds (EX1). An explanation is a proposition which was previously unknown to the agent (EX4), and expresses a minimal event (EX3) which is an actual cause of the explanandum in every world in the agent's epistemic state in which it is true (EX2).

In addition to incorporating consumer-dependence as in Gärdenfors's proposal, and involving causality in a thoroughgoing way, the HP account enjoys the virtue of formal precision. At the same time, Halpern and Pearl are not entirely clear on how the various formal components relate to actual practices of giving and receiving explanations. Part of this unclarity arises when we try to assess potential counterexamples to the account. For instance, it has been observed in the empirical literature that an explanans is often something that the consumer already knows, in evident violation of EX4 (Kirfel, Icard, & Gerstenberg, 2022). Consider the following, adapted from Faye (2007):

**Example 1** (Roof Replacement)**.** A house catches fire. Two things caused the house to catch fire: the fact that the roof was thatched and the fact that it hadn't rained recently. Bob asks, "why did the house catch fire?". Suppose that Bob knows the house had a thatched roof and that it hadn't rained recently (Bob knows the state of the world), but doesn't know enough about how fires start to know if either of these factors are the sorts of factors that cause fires (Bob is uncertain about the world's causal structure).

Suppose that in fact the house caught fire because the roof was thatched. Most people will have the intuition that a good answer to Bob's question is something that he already knows, namely, "it had a thatched roof". Not only is this a possible answer, it seems the best answer. Thus, on the face of it, Example 1 is a straightforward counterexample to EX4. At the same time, one could argue that the "surface form" of this answer to Bob's question – which only mentions the thatched roof – conveys new causal information, and it is this implicit causal information that is the real explanans. For instance, Bob might learn that thatched roofs *cause* wildfire spread.

In light of examples like this, Halpern and Pearl could perhaps be interpreted as proposing an account of what a person learns from an explanation in context. On this reading, however, it becomes somewhat unclear what role the conditions other than EX2 (actual causation) are to play. For instance, why should we be concerned with minimality (EX3), which can be more easily motivated by appeal to communicative pressure? If anything, a good explanation ought to bear *more* rather than less inferential fruit. Absent EX3 and EX4, the Halpern-Pearl account just says that the explanandum should be true and the explanans should pick out an actual cause.

We suggest that, like the formal accounts that came before, the HP definition marks important progress on analysing ex-

planation. However, the substantive components of their theory – to the extent that these further conditions are apt in the first place – should instead emerge, perhaps in a defeasible and graded way, from a more fundamental analysis of what explanations do in conversational context.

**HP on Explanatory Goodness** Before turning to our proposal, it is worth mentioning how a framework like this can compare different explanations. Halpern and Pearl consider a probability distribution Prior on $\mathcal{K}$, representing the agent's uncertainty. Given explanandum FACT, let $\mathcal{K}_{\mathbf{X}=\mathbf{x}}$ be the largest subset $\mathcal{K}^*$ of $\mathcal{K}$ such that $\mathbf{X}=\mathbf{x}$ is an explanation of Fact relative to $\mathcal{K}^*$. The *goodness of a potential explanation* $\mathbf{X}=\mathbf{x}$ for FACT is given by $\mathsf{Prior}(\mathcal{K}_{\mathbf{X}=\mathbf{x}} \mid \mathbf{X}=\mathbf{x})$; that is, by the probability that $\mathbf{X}=\mathbf{x}$ is an explanation of FACT (according to Def. 2), conditional on its being true.

Note that, by this definition, an explanation could have arbitrarily high goodness (including 1) with arbitrarily low probability. Given the resources available in the HP framework, it is unclear whether there are better ways of combining these formal primitives to capture explanatory goodness. On the communication-theoretic account that we favour, capturing explanatory goodness will require new theoretical primitives: explicit incorporation of a speaker/producer, and explication of listener/consumer interests.

## A Communication-Theoretic Account

As for Halpern and Pearl, we assume that an agent's uncertainty over $\mathcal{K}$ is represented by a probability distribution Prior. We will suppose further that this agent has asked (perhaps explicitly) the question "why FACT?", where FACT is something he knows. Call this agent the *listener*. Beyond the setup from Gärdenfors, Halpern and Pearl, we assume that a second agent is involved: call this agent the *speaker*.

To model the interaction between these two agents, we draw upon a body of work at the intersection of linguistics and psychology intended to capture pragmatics of conversation. This work, sometimes referred to as the *rational speech acts* (or RSA) framework, can be seen as a formalization of Gricean ideas from the philosophy of language (Degen, 2023; Goodman & Frank, 2016; Frank & Goodman, 2012; Sumers, Ho, Griffiths, & Hawkins, 2023). It involves a hierarchy of conversational agent models, typically beginning with a "literal" listener, who interprets messages according to their semantic content, then ascending to a pragmatic speaker, who selects a message taking into account how the message will be interpreted by the literal listener. Last comes a pragmatic listener who interprets the pragmatic speaker's utterance as a "rational speech act". We follow this presentation here.

### The Literal Listener

Imagine a literal listener – named *L0* – who has posed the question, "why FACT?", and now hears a response, "because $\mathbf{X}=\mathbf{x}$". What might *L0* do with this response?

Intuitively, what *L0* learns – that is, how Prior should be updated – will depend on what the message "FACT because $\mathbf{X}=\mathbf{x}$" says about the causal facts. Let us suppose that "FACT because $\mathbf{X}=\mathbf{x}$" is true of some model-context pairs $(\mathcal{M},\mathbf{u})$, and false at others. In particular, we will suppose that this statement is true just in case $\mathbf{X}=\mathbf{x}$ is an actual cause of FACT. For the purposes of this paper, virtually any extant account of actual cause from the literature will be suitable here.

Formally, we thus have a message $m$ (e.g., "FACT because $\mathbf{X}=\mathbf{x}$") that has a semantic value, $[\![m]\!] = \{(\mathcal{M},\mathbf{u}) \in \mathcal{K} \mid \mathcal{M},\mathbf{u} \models m\}$, where $\mathcal{M},\mathbf{u} \models m$ means that $m$ is true of the pair $(\mathcal{M},\mathbf{u})$. Following the RSA framework, we then suppose that *L0* simply updates his uncertainty toward a "posterior" distribution $P_{L0}$ as: $P_{L0}(\mathcal{M},\mathbf{u} \mid m) \propto \mathsf{Prior}(\mathcal{M},\mathbf{u}) \cdot \mathbf{1}_{\mathcal{M},\mathbf{u}\models m}$. In words, message $m$ leads *L0* to assign probability 1 to $m$ being true, and to redistribute probability mass among the remaining possibilities in proportion to his prior beliefs.

### The Pragmatic Speaker

Imagine that a speaker is aware of her addressee, *L0*, and how *L0* will respond to various possible responses she could give to *L0*'s "why?" question. How should the speaker – let us call her *S* – choose a response?

**Useful Utterances** While much work in the RSA framework has focused on what we might call "pure information exchange", a natural thought – explored in recent work by Sumers et al. (2023) – is that a speaker will convey information that helps the listener achieve his goals. Let us suppose that the listener has a decision problem, characterized by a pair $(\mathcal{A},\mathcal{R})$, where $\mathcal{A}$ is a set of actions and $\mathcal{R}: \mathcal{A} \times \mathcal{K} \to \mathbb{R}$ is a *reward function*, specifying how good some action $a$ is in possible causal situation $(\mathcal{M},\mathbf{u})$. That is, the agent will choose an action $a$ and receive some scalar reward $\mathcal{R}(a,\mathcal{M},\mathbf{u})$. If he knew which possibility in $\mathcal{K}$ in fact obtained, then he would simply choose the action that returns the highest reward in that situation. But recall the listener has some uncertainty, represented by Prior.

A cooperative speaker will thus produce an utterance that leads the listener toward better actions, by updating the listener's uncertainty. Recall that a message $m$ will lead *L0* to update Prior to a posterior distribution $P_{L0}(- \mid m)$. We assume that the listener will then (approximately) maximize expected reward with respect to $P_{L0}$. Specifically, we assume that the probability with which *L0* chooses action $a$ given message $m$ is such that $\pi_{L0}(a \mid m)$ is proportional to

$$\exp\left(\beta_L \cdot \left[\sum_{(\mathcal{M},\mathbf{u})\in\mathcal{K}} P_{L0}(\mathcal{M},\mathbf{u} \mid m) \cdot R(a,\mathcal{M},\mathbf{u})\right]\right),$$

where $\beta_L > 0$ is a "rationality parameter", measuring how close the agent is to maximizing expected utility. This choice follows a large body of work in psychology, economics, and computer science modelling agents as approximate expected utility maximizers (see, e.g., Luce, 1963).

If *L0* is employing decision policy $\pi_{L0}$, then the reward that *S* can expect of *L0* is given by

$$U_S(m,\mathcal{M},\mathbf{u}) \quad = \quad \sum_{a\in\mathcal{A}} \pi_{L0}(a \mid m) \cdot R(a,\mathcal{M},\mathbf{u}).$$

$S$ ought to choose her utterance $m$ – in our setting, her answer to the listener's "why?" question – in a way that (approximately) maximizes this anticipated utility $U_S$:

$$P_S(m \mid \mathcal{M}, \mathbf{u}) \quad \propto \quad \exp\Big(\beta_S \cdot U_S(m, \mathcal{M}, \mathbf{u})\Big).$$

Often, $\beta_L$ and $\beta_S$ can be simply set to $\infty$ (positive infinity), meaning that agents are maximizing expected utility.

**Production and Processing Costs** Part of what is interesting about explanations is that they must be constructed and interpreted by resource-limited humans. One can imagine several ways of incorporating such resource limitations. For instance, $S$ may be aware of the possibility that $L0$ could misinterpret a message, or indeed that $S$ herself may err in her production of the message. In this vein we could incorporate assumptions $S$ could make about the *channel capacity* for messages from $S$ to $L0$ (Gibson et al., 2019).

An alternative, which we will adopt here, is that both production and processing of a message $m$ come with some measurable cost. We put these two sources of cost together into a single $\mathsf{Cost}(m)$. The speaker probabilities now become

$$P_S(m \mid \mathcal{M}, \mathbf{u}) \quad \propto \quad \exp\left(\beta_S \cdot \Big[U_S(m \mid \mathcal{M}, \mathbf{u}) - \mathsf{Cost}(m)\Big]\right).$$

Our model thus assumes that the speaker and listener will produce and interpret each message correctly; they may just suffer cost in doing so, due to length, obscurity, complexity, etc.

## What is the Listener's Decision Problem?

The idea that explanation can depend on interests and goals of an listener is a perennial theme across pragmatic approaches to explanation (Van Fraassen, 1980; Potochnik, 2017; De Regt, 2017; Lombrozo & Liquin, 2023). As our aim is to offer a precise framework, we will need to be somewhat concrete about what sorts of decision problems $(\mathcal{A}, \mathcal{R})$ agents might face. At least in some cases, context will render it common knowledge that the listener has asked a "why?" question so as to inform a particular future choice.

In other cases, though, the speaker may be uncertain about which of a number of decision problems the listener might face. Thus, we could also imagine that the decision problem $(\mathcal{A}, \mathcal{R})$ decomposes into a collection of decision problems $(\mathcal{A}_i, \mathcal{R}_i)$, each with a weight $w_i$. Where an action $a$ is now a vector $(\ldots, a_i \ldots)$, specifying choices for all the decision problems, the total reward would be given by the sum $\mathcal{R}(a, \mathcal{M}, \mathbf{u}) = \sum_i w_i \mathcal{R}_i(a_i, \mathcal{M}, \mathbf{u})$.

**The Manipulation Game** Suppose, however, that the speaker cannot enumerate a specific list of possible decision problems, together with their weights. Instead, she might want to impart causal information that broadly promises to be useful. A fundamental intuition from the interventionist tradition in the philosophy of causation and explanation is that general possibilities for *manipulation* and *control* are of central importance (Woodward, 2003; Kirfel et al., 2024). We formalize what it might mean to have relatively broad capacity for manipulation and control with what we call a "manipulation game".

In the simplest case, we imagine the listener having asked, "why FACT?" suggests that causal information about FACT is somehow relevant to the decision problems he faces. As a general proxy for whatever those decision problems might be, imagine the following game. The listener is presented with some alternative way the world might have been (that is, different assignments to the exogenous variables in his knowledge state), with probability proportional to their prior likelihood. For each such possibility, the agent must choose some endogenous variable to intervene upon, with the aim of changing the truth value of FACT. In each such case, the agent wins a point just in case he successfully manipulates FACT in that situation. More formally:

**Definition 3** (Manipulation Game)**.** A manipulation game is a decision problem $(\mathcal{A}, \mathcal{R})$, where:

- $\mathcal{A}$ is the set of all endogenous variables other than those appearing in FACT.
- $\mathcal{R} : \mathcal{A} \times \mathcal{K} \to \mathbb{R}$ is defined, for a given endogenous variable $X \in \mathcal{A}$ and possibility $(\mathcal{M}, \mathbf{u}) \in \mathcal{K}$, as $\mathcal{R}(X, \mathcal{M}, \mathbf{u}) =$

$$\sum_{\mathbf{u}' \in \mathsf{Val}(\mathbf{U})} P(\mathbf{u}') \cdot \mathsf{Manipulates}(X, \mathsf{FACT} \mid \mathcal{M}, \mathbf{u}')$$

  where $\mathsf{Manipulates}(X, \mathsf{FACT} \mid \mathcal{M}, \mathbf{u}')$ is a binary-valued function that takes 1 iff there exists $x \in \mathsf{Val}(X)$ such that either $\mathcal{M}, \mathbf{u}' \models \mathsf{FACT} \wedge [X = x] \neg \mathsf{FACT}$, or $\mathcal{M}, \mathbf{u}' \models \neg \mathsf{FACT} \wedge [X = x] \mathsf{FACT}$.

Less formally, the listener submits an endogenous variable to intervene on. For each possible context, he receives a score according to whether manipulating the value of the variable manipulates the value of FACT, weighted by the probability of the context. Note in particular that $\mathcal{R}(X, \mathcal{M}, \mathbf{u})$ is not sensitive to the actual context $\mathbf{u}$.

This reward function is closely related to several models of *causal strength* in the psychological literature. When all the variables are binary it coincides with (a causal version of) the so-called $\Delta P$ measure (see Jenkins & Ward, 1965; Cheng & Novick, 1992). Replacing $P(\mathbf{u}')$ with the sampling procedure in Lucas and Kemp (2015) gives the so-called *counterfactual effect size model* of (Quillien & Lucas, 2023).

## The Pragmatic Listener and Explanatory Goodness

Having described the literal listener, the pragmatic speaker, and the range of possible decision problems the speaker could entertain for the listener, we are now ready to present the final component, a pragmatic listener. This last agent-type will update his beliefs not directly based on the content of the utterance, but based on the fact that the speaker made this particular utterance, knowing what she does about the (literal) listener and the decision problems confronting him.

This pragmatic listener – who we call $L$ – updates his prior in a way that is closely analogous to $L0$, the literal listener:

$P_L(\mathcal{M},\mathbf{u}\,|\,m) \propto \mathsf{Prior}(\mathcal{M},\mathbf{u}) \cdot P_S(m\,|\,\mathcal{M},\mathbf{u})$. The only difference is in the second term, with the indicator function on the truth of $m$ (in situation $\mathcal{M},\mathbf{u}$) replaced by the probability that $S$ would utter $m$. This in turn supplies us with an updated agent policy $\pi_L(a\,|\,m)$, again identical to the expression for $\pi_{L0}$, except that we replace $P_{L0}$ with $P_L$.

With this much we are now ready to offer our proposal about explanatory "goodness". In rather stark contrast to the HP definition, we submit that this can be measured as:

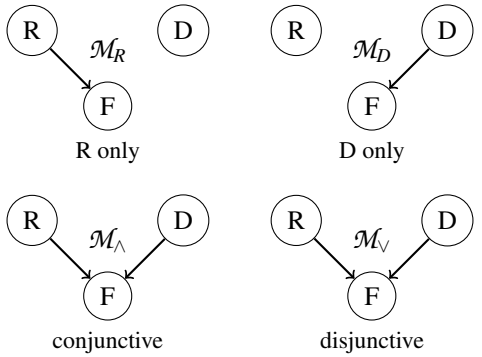
$$\mathsf{Goodness}(m,\mathcal{M},\mathbf{u}) \quad = \quad \sum_{a\in\mathcal{A}} \pi_L(a\,|\,m)\cdot R(a,\mathcal{M},\mathbf{u}) - \sum_{a\in\mathcal{A}} \pi_{\mathsf{Prior}}(a\,|\,m)\cdot R(a,\mathcal{M},\mathbf{u}).$$

In words, $m$ is a good explanation to the extent that it helps the listener achieve his goals. So while actual causation does play a fundamental role, it does so only instrumentally in the way it facilitates success at a range of downstream tasks.

## Explaining Explanation

We submit that this basic model already accounts for many, if not most, of the "explanatory virtues" that have been discussed in the literature (see De Regt, 2017), in addition to key experimental patterns on explanatory judgment. We discuss only a few examples here, leaving others for future work.

Developing Example 1, we could model this situation with three binary endogenous variables $R,D,F$, and two binary exogenous variables $U_R,U_D$. $F$ represents whether the house catches fire; $R$ whether its roof is thatched; and $D$ whether there was a recent drought. Recall that Bob knows $R,D,F=1$, but is uncertain about the causal structure; he believes it is possible that $R=1$ but not $D=1$ causes $F=1$ (structure $\mathcal{M}_R$), that $D=1$ but not $R=1$ causes $F=1$ (structure $\mathcal{M}_D$), that $F=1$ iff $R=1$ *and* $D=1$ (a *conjunctive* structure) or that $F=1$ iff $R=1$ *or* $D=1$ (a *disjunctive* structure). The models appear below, exogenous variables omitted for readability.




Let us suppose the speaker has two utterances available to her, $R=1$ (interpreted as asserting that the roof's being thatched was a cause of the fire) and $D=1$ (the drought was a cause of the fire). All that remains is to specify a decision problem.

### A specific decision problem: distinguishing informativeness and usefulness

Suppose first that Bob's own roof is thatched, and he's wondering whether or not to replace it. He would prefer to replace his roof ($a_{\text{replace}}$) if thatched roofs cause fires, but otherwise he'd prefer not to pay the expense ($a_{\text{leave}}$). We represent this with the pay-off matrix below. Suppose that Prior is uniform

| | $\mathcal{M}_R$ | $\mathcal{M}_D$ | $\mathcal{M}_\wedge$ | $\mathcal{M}_\vee$ |
|---|---|---|---|---|
| $a_{\text{replace}}$ | 0 | 0 | 0 | 0 |
| $a_{\text{leave}}$ | $-1$ | 1 | $-1$ | $-1$ |

over $\mathcal{K}$, and focus on the case where the actual causal structure is $\mathcal{M}_\wedge$. Then although $P_{L0}(\mathcal{M}_\wedge\,|\,R=1) = P_{L0}(\mathcal{M}_\wedge\,|\,D=1)$ (i.e. citing the thatched roof and drought as causes are equally *informative* to Bob as to the true causal structure), we have $U_S(R=1,\mathcal{M}_\wedge) > U_S(D=1,\mathcal{M}_\wedge)$, because it is more *useful* to Bob to learn that the roof's being thatched was a cause (it allows him to make the sensible decision to replace his roof).

### Uncertainty over two decision problems: prioritising unknown causes

Suppose next that the speaker thinks it's possible that Bob is considering another decision problem, namely whether or not to move to an area which is unaffected by drought. He'd rather not move, but would prefer to do so if droughts cause fires. We represent this with the pay-off matrix below. Suppose that the speaker considers it equally likely that

| | $\mathcal{M}_R$ | $\mathcal{M}_D$ | $\mathcal{M}_\wedge$ | $\mathcal{M}_\vee$ |
|---|---|---|---|---|
| $a_{\text{move}}$ | 0 | 0 | 0 | 0 |
| $a_{\text{stay}}$ | 1 | $-1$ | $-1$ | $-1$ |

Bob faces this decision problem as the previous one, and wants to maximise his performance on a weighted average of the two. When Prior is uniform over $\mathcal{K}$, we will have $U_S(R=1,\mathcal{M}_\wedge) = U_S(D=1,\mathcal{M}_\wedge)$ as is easily seen. Suppose, though, that Bob knows that drought is a cause of the fire (i.e. that Prior is instead uniform over $\{\mathcal{M}_D,\mathcal{M}_\wedge,\mathcal{M}_\vee\}$). Then we will have $\mathsf{Goodness}(R=1,\mathcal{M}_\wedge) > \mathsf{Goodness}(D=1,\mathcal{M}_\wedge)$. So our model accounts for the intuition that it is better to cite causes which are unknown to the listener, without simply stipulating this (contra, e.g., the HP model in Def. 2).

### Manipulation, causal selection and normality

Notably, our model accounts for recent empirical work exploring the affects of "normality" on the probability of citing a cause. In what follows, we assume that $P(U_R) < 0.5 < P(U_D)$; that is, droughts are relatively common, while the roof's being thatched is a statistically "abnormal" event. Suppose that Alice is uncertain about the precise decision problem Bob faces, and wants to give him information which will be useful for a variety of decision problems. We model this

by supposing Bob is facing a manipulation game (Def. 3) with the following payoff matrix:

| | $\mathcal{M}_R$ | $\mathcal{M}_D$ | $\mathcal{M}_\wedge$ | $\mathcal{M}_\vee$ |
|---|---|---|---|---|
| $R$ | 1 | 0 | $P(U_D)$ | $1-P(U_D)$ |
| $D$ | 0 | 1 | $P(U_R)$ | $1-P(U_R)$ |

**(Ab)normal Causal Selection** In conjunctive causal structures people prefer to cite abnormal causes over normal causes, and vice-versa in disjunctive structures (Gerstenberg & Icard, 2020; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Quillien & Lucas, 2023). Crucially, our model captures both patterns. On the manipulation game, we have $\mathcal{R}(R,\mathcal{M}_\wedge) > \mathcal{R}(D,\mathcal{M}_\wedge)$ and $\mathcal{R}(R,\mathcal{M}_\vee) < \mathcal{R}(D,\mathcal{M}_\vee)$. From these two facts it follows that $U_S(R=1,\mathcal{M}_\wedge) > U_S(D=1,\mathcal{M}_\wedge)$, and also that $U_S(R=1,\mathcal{M}_\vee) < U_S(D=1,\mathcal{M}_\vee)$.

**Inference from Normality** Aside from production, recent work also shows that a listener can infer a great deal from a speaker's choice of which cause to cite. For instance, people can infer whether a structure is disjunctive or conjunctive (when they know which cause is normal) or infer which cause is normal (when they know the causal structure; Kirfel et al., 2022). Again, our model captures this, via the 'pragmatic listener' $L$. For example, suppose Alice cites the abnormal cause (the roof's being thatched, $R=1$). Then, since we have $P_S(R=1,\mathcal{M}_\wedge) > P_S(R=1,\mathcal{M}_\vee)$, we have

$$P_L(\mathcal{M}_\wedge \,|\, R=1) > P_L(\mathcal{M}_\vee \,|\, R=1).$$

In other words, the pragmatic listener is able to infer that the structure is conjunctive rather than disjunctive, since he takes into account the fact that the speaker will prefer to cite abnormal causes when the structure is conjunctive (and normal causes when it is disjunctive). Similarly, we can see that the pragmatic listener is able to infer that the structure is disjunctive when the speaker cites the normal cause. It is worth emphasizing again that patterns like these appear to violate HP's fourth postulate, EX4. While it is often appropriate to cite unknown factors – as in the scenario above with two decision problems – this cannot be a hard requirement on explanation.

**Minimality** HP require (EX3, Def 2) that explanations are minimal. But there are many cases in which including additional detail makes an explanation better (Salmon, 2001; Zemla, Sloman, Bechlivanidis, & Lagnado, 2023). Our model accounts for this; a soft preference for minimality instead emerges from communicative pressures.

## Other Patterns and Virtues

Without building in any of these specific patterns, we find that they naturally emerge from the way that speakers and listeners interact around actual causation. Many additional patterns emerge from this model. For instance, good explanations

1. identify explanatory relationships that are invariant across background conditions (Lewis, 1986; Woodward, 2006; Ylikoski & Kuorikoski, 2010);

2. aim at the right level of abstraction (Yablo, 1992; Strevens, 2008; Woodward, 2021);

3. achieve some type of generality, unifying seemingly diverse phenomena (Friedman, 1974; Kitcher, 1981);

and so on. Space prohibits verifying each of these further patterns; we leave this to future work.

## Conclusion

What makes a good answer to a "why?" question? Our proposal employs a variant of the rational speech act model, as extended by Sumers et al. (2023). We have a cascade of agent-types, with a literal listener at the bottom, followed by a pragmatic speaker thinking about her likely effect on the literal listener, and finally a pragmatic listener thinking about what prompted the speaker to say what she did. This is all grounded in the speaker's motivation to help the listener achieve some goal. What is special about explanation is twofold. First and most fundamentally, we assume that "why?" questions are requests for causal information, and when asking about singular events, for information about actual causes. This assumption is built into the semantics of "because" statements. Second, listeners are often interested in fairly general issues of manipulation and control, an assumption that is formalized via the manipulation game.

The framework can be seen as a natural, formal combination of rather general pragmatic reasoning on the one hand, and the interventionist idea that explanation is about asking and answering what-if-things-had-been-different questions on the other. These two ingredients, combined in the simple way we have done here, are already enough to account for some of the most striking features of (good) explanation that have been discussed in the literature.

Although we claim that many important patterns emerge nearly "for free" on this account, there remains significant work to flesh out further details. As an example, we noted that virtually any extant account of actual causation could be slotted into our framework. There may nonetheless be examples that depend on this choice; and if there are not, it would be desirable to extract a simpler analysis of this notoriously bewildering concept (see, e.g., Glymour et al., 2010; Hitchcock, 2017). Similarly, there will be contexts (e.g. scientific enquiry) in which the listener asks "why" questions for which he will never face a concrete decision problem; extending the framework to these domains requires philosophical work. Finally, it has been argued that the human drive toward explanation serves several auxiliary psychological functions, such as learning, generalization, etc. (Lombrozo, 2006). Demonstrating how these functions emerge through "explaining to oneself" would lend further support to the present framework and its ability to explain explanation.

An extended version of this paper appears here.

## References


Achinstein, P. (1983). *The nature of explanation*. Oxford University Press.

Bareinboim, E., Correa, J., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), *Probabilistic and causal inference: The works of Judea Pearl* (p. 509-556). ACM Books.

Bromberger, S. (1965). An Approach to Explanation. In R. J. Butler (Ed.), *Analytic Philosophy, 2nd Edition.* Blackwell.

Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, *99*, 365-382.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

De Regt, H. W. (2017). *Understanding scientific understanding*. Oxford University Press.

Douven, I. (2022). *The art of abduction*. MIT Press.

Dowe, P. (1992). Wesley salmon's process theory of causality and the conserved quantity theory. *Philosophy of Science*, *59*(2), 195–216.

Faye, J. (2007). The pragmatic-rhetorical theory of explanation. In J. Persson & P. Ylikoski (Eds.), *Rethinking Explanation. Series: Boston Studies in the Philosophy of Science Vol. 252*. (pp. 43–68). Springer Verlag.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Friedman, M. (1974). Explanation and Scientific Understanding. *The Journal of Philosophy*, *71*(1), 5–19. doi: 10.2307/2024924

Gallow, D. (2021). A model-invariant theory of causation. *The Philosophical Review*, *130*(1), 45–96.

Gärdenfors, P. (1980). A Pragmatic Approach to Explanations. *Philosophy of Science*, *47*(3), 404–423.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states*. MIT Press.

Gärdenfors, P. (1990, September). An epistemic analysis of explanations and causal beliefs. *Topoi*, *9*(2), 109–124. doi: 10.1007/BF00135892

Gerstenberg, T., & Icard, T. F. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

Gibson, E., Futrell, R., Piantadosi, S. P., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How efficiency shapes human language. *Trends in Cognitive Sciences*, *23*(5), 389–407.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., ... Zhang, J. (2010). Actual causation: a stone soup essay. *Synthese*, *175*(2), 169–192.

Goddu, M. K., & Gopnik, A. (2024). The development of human causal learning and reasoning. *Nature Reviews Psychology*, 1–21.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Halpern, J. Y., & Pearl, J. (2005a). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843-887.

Halpern, J. Y., & Pearl, J. (2005b). Causes and Explanations: A Structural-Model Approach. Part II: Explanations. *The British Journal for the Philosophy of Science*, *56*(4), 889–911.

Hempel, C. G. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (No. 1). The Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, *15*(2), 135–175.

Henne, P., Niemi, L., Pinillos, A., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157-164.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*(6), 273-299.

Hitchcock, C. (2017). Actual causation: What's the use? In *Making a difference: Essays on the philosophy of causation.* Oxford University Press.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017, April). Normality and actual causal strength. *Cognition*, *161*, 80–93. doi: 10.1016/j.cognition.2017.01.010

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*(1), 1-17.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*(Volume 57, 2006), 227-254.

Kirfel, L., Harding, J., Shin, J., Xin, C., Icard, T., & Gerstenberg, T. (2024). Do as I explain: Explanations communicate optimal interventions. In L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Kirfel, L., Icard, T., & Gerstenberg, T. (2022). Inference from explanation. *Journal of Experimental Psychology: General*, *151*(7), 1481–1501. doi: 10.1037/xge0001151

Kitcher, P. (1981). Explanatory unification. *Philosophy of Science*, *48*(4), 507–531.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D. A., & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196–209.

Lewis, D. (1986). Causal explanation. In D. Lewis (Ed.), *Philosophical Papers Vol. Ii* (Vol. 2, pp. 214–240). Oxford University Press.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470.

Lombrozo, T., & Liquin, E. G. (2023). Explanation is effective because it is selective. *Current Directions in Psychological Science*, *32*(3), 212-219.

Lucas, C. G., & Kemp, C. (2015). An Improved Probabilistic Account of Counterfactual Reasoning. *Psychological Review*, *122*(4), 700–734. doi: 10.1037/a0039655

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (p. 103-189). Wiley.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669-710.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. Cambridge, MA: The MIT Press.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Quillien, T., & Lucas, C. G. (2023). Counterfactuals and the logic of causal selection. *Psychological Review*.

Rottman, B. M., & Hastie, R. (2016). Do people reason rationally about causally related events? markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134.

Salmon, W. C. (1971). *Statistical explanation & statistical relevance*. University of Pittsburgh Press.

Salmon, W. C. (1998). *Causality and explanation* (Vol. 52). Oxford University Press.

Salmon, W. C. (2001). Reflections of a bashful Bayesian: a reply to Peter Lipton. In *Explanation* (p. 121–136). Springer.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223–247.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Springer Lecture Notes in Statistics.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Harvard University Press.

Sumers, T. R., Ho, M. K., Griffiths, T. L., & Hawkins, R. D. (2023). Reconciling truthfulness and relevance as epistemic and decision-theoretic utility. *Psychological Review*.

Van Fraassen, B. C. (1977). The Pragmatics of Explanation. *American Philosophical Quarterly*, *14*(2), 143–150.

Van Fraassen, B. C. (1980). *The scientific image* (No. 4). Oxford University Press.

Waldmann, M. R. (Ed.). (2017). *The oxford handbook of causal reasoning*. Oxford University Press.

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J. (2006). Sensitive and Insensitive Causation. *The Philosophical Review*, *115*(1), 1–50.

Woodward, J. (2021, January). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese*, *198*(1), 237–265.

Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part I: A counterfactual account. *Noûs*, *37*(1), 1-24.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*(2), 245–280.

Ylikoski, P. (2007). The Idea of Contrastive Explanandum. In J. Persson & P. Ylikoski (Eds.), *Rethinking Explanation* (pp. 27–42). Springer.

Ylikoski, P., & Kuorikoski, J. (2010). Dissecting explanatory power. *Philosophical Studies*, *148*(2), 201–219.

Zemla, J. C., Sloman, S. A., Bechlivanidis, C., & Lagnado, D. A. (2023). Not so simple! causal mechanisms increase preference for complex explanations. *Cognition*, *239*, 105551.