Annals of Mathematics and Philosophy
***vol.* 3, *no.* 1**, 2025
https://doi.org/10.67678/ntxnceqn

$M \times \Phi$

Explanations of Mathematical Statements

Joseph Y. Halpern[(*)]

Abstract. A definition of what counts as an explanation of a mathematical statement, and when one explanation is better than another, is given. Since all mathematical facts must be true in all causal models, and hence known by an agent, mathematical facts cannot be part of an explanation (under the standard notion of explanation). This problem is solved using impossible possible worlds.

Keywords. mathematical explanation, partial explanation, causality, impossible possible worlds

[(*)]Supported in part by NSF grants IIS-178108 and IIS-1703846, MURI grant W911NF-19-1-0217, and ARO grant W911NF-22-1-0061. I thank Jordan Ellenberg for asking the question that motivated this paper, and for useful conversations on the subject of explanations of mathematical results, and the reviewers of this paper for their useful comments.

§ 1. — Introduction.

Consider two mathematical questions like "Why is 4373 the sum of two squares?" or "For $f(x) = x^{11} - 6x^{10} + 11x^9 - 17x^8 + 22x^7 - 5x^6 + 10x^5 + x^9 - 2x^3 - x + 2$, why is $f(2) = 0$?". An explanation for the first question might simply be a demonstration: $4373 = 3844 + 529 = 62^2 + 23^2$. An arguably better explanation, at least for those who know Fermat's two-squares theorem, which says that an odd prime is congruent to 1 mod 4 (i.e., has a remainder of 1 when divided by 4) if and only if it is the sum of two squares, is the observation that 4373 is 1 mod 4 (since $4373 = 4 \times 1093 + 1$) and is prime. Similarly, an explanation for the second question is just the observation that (after some laborious multiplication and addition)

$$2^{11} - 6 \times 2^{10} + 11 \times 2^9 - 17^8 + 22 \times 2^7 - 5 \times 26 + 10 \times 2^5 + 2^9 - 2 \times 2^3 - 2 + 2 = 0; \quad (1)$$

a better explanation (at least for many) is the observation that

$$f(x) = (x-2)(x^{10} - 4x^9 + 3x^8 = 11x^7 - 5x^5 + x^3 - 1). \quad (2)$$

What makes these explanations? Why is one explanation better than another? There have been many definitions of explanation proposed in the literature. Hempel's (1965) *deductive-nomological* model does a good job of accounting for why these count as explanations. In this model, an explanation consists of a "law of nature" and some additional facts that together imply the *explanandum* (the fact to be explained). In the first example, the law of nature is Fermat's two-squares theorem; the additional facts are the observation that 173 is a prime and that it is congruent to 1 mod 4. Similarly, in the second example, the law of nature is the fact that if $(x - m)$ is a factor of a polynomial f whose coefficients are integers (i.e., $f(x) = (x-2)g(x)$, where g is a polynomial whose coefficients are integers), then $f(2) = 0$, together with (2).

As is well known, there are difficulties with Hempel's account of explanation. For one thing, it does not take causality into account; for another, it does not take into account the well-known observation that what counts as an explanation is relative to what an agent knows (Gärdenfors 1988, Salmon 1984). As Gärdenfors (1988) observes, an agent seeking an explanation of why Mr. Johansson has been taken ill with lung cancer will not consider the fact that he worked for years in asbestos manufacturing an explanation if he already knew this fact. Finally, this definition does not give us a way to say that one explanation is better than another.

In this paper, I focus on the definition of (causal) explanation given by Halpern and Pearl (Halpern 2016, Halpern and Pearl

2005b) (HP from now on). It has been shown to do quite well on many problematic examples for which other definitions of explanations have difficulty. The definition starts with a definition of causality in a causal model. For our purposes, a causal model can be described by an acyclic graph. Each node in the graph other than the roots of the graph is associated with an equation. These equations can be viewed as encoding the steps of a proof. In a causal model, we can talk about one event being a cause of another. Explanation is then defined relative to an agent's *epistemic state*, which, roughly speaking, consists of a set of casual models and a probability on them. The epistemic state can be thought of as describing the agent's beliefs about how causality works in the world; an agent is said to know a fact if the fact has probability 1 according to his epistemic state. Following Gärdenfors, we would expect an explanation to be something that the agent did not already know.

But now there seems to be a problem: all mathematical facts must be true in all causal models, and hence must be known by an agent. That means mathematical facts cannot be part of an explanation. On the other hand, how can we give an explanation of a mathematical statement without invoking facts of mathematics?

In this note, I sketch a solution to this problem using the HP definition of explanation; I expect that the idea might well apply to other definitions as well. The solution takes as its point of departure the notion of "impossible" possible worlds. This idea has a long history in epistemic logic (Cresswell 1970, Cresswell 1972, Cresswell 1973, Hintikka 1975, Kripke 1965, Rantala 1982). To understand it, recall that Hintikka (1962) assumed that an agent considered a number of worlds possible, and took the statement "Agent a knows φ" to be true if φ was true in all the worlds that an agent considered possible. So one way to model the fact that a knows it's sunny in Ithaca and doesn't know whether it's sunny in Berkeley is by saying a considers two worlds possible; in one, it's sunny in both Ithaca and Berkeley, while in the other, it's sunny in Ithaca and raining in Berkeley. With this viewpoint, "possibility" is the dual of knowledge. Agent a considers φ possible if a does not know not φ, which is the case if a considers at least one world possible where φ is true.

This approach also runs into trouble with mathematical statements. Suppose that agent a is presented with a 200-digit number n. It seems reasonable to say a doesn't know whether n is prime; a considers it possible both that n is prime and that n isn't prime.

But suppose that n is in fact prime. That would mean that the world that a considers possible where n is not prime is inconsistent with basic number theory. The "impossible" possible worlds approach referred to above allows agent a to consider such worlds possible.

Here I show that the analogous approach, when applied to causal models, allows us to deal with the explanation of mathematical statements. Specifically, in the case of the sum of squares, the epistemic state would include causal models where 4373 is not a prime (and/or causal models where it is not congruent to 1 mod 4); in the case of the polynomial, the epistemic state would include causal models where $x - 2$ is not a factor of f. This approach will also allow us to say that one explanation of a mathematical statement is better than another.

It might be argued that what I am doing here is not really applicable to explanations of mathematical statements, and that it is inappropriate to view proofs of mathematical statement as "causal" explanations, or to view various mathematical facts as "causes" of other mathematical facts. To me, there is no real conceptual difference between viewing the fact that 4373 is congruent to 1 mod 4 as an explanation of the fact that it is the sum of two squares and viewing the fact that a child spent a year learning remotely due to COVID as an explanation of his poor performance the following year. Moreover, in natural language, we often make statements "the reason that the theorem is true is because ...", followed by a proof of the theorem. The use of the word "because" suggests (at least to me) that people think of proofs as providing causal explanations (even though the mathematical facts being proved are not themselves causal).

Given the (quite extensive) literature on mathematics and explanation (see (Pincock 2023) for an overview and references), it is perhaps useful to compare more generally what I do in this paper to that literature. Most of that literature views "mathematical explanation" as a special type of explanation, and tries to explain what it makes it special. Pincock analyzes various approaches to explanations of mathematics in terms of five principles. The fourth one specifically addresses this issue; it says "There is a special way that mathematics may appear in a scientific explanation that makes it a genuine mathematical explanation". My goal is not to characterize mathematical explanations, nor to carefully distinguish mathematical explanations from other types of explanations. Indeed, I'm not sure that such a distinction exists. Rather, I try to

show that mathematical explanations can be viewed as a special case of the HP approach to explanation. As a result, I suspect that my approach would not satisfy Pincock's fourth principle. (I am not sure, because I'm not sure what would count as a "special way" that mathematics appears in the explanation.) That said, I hope that even those who do view mathematical explanation as a distinct form of explanation will find the exercise of showing that they can be viewed as an instance of the HP definition of interest.

It is also worth noting that the idea of using impossible possible worlds when defining mathematical proofs is not new. Specifically, Baron, Colyvan, and Ripley (2017) use impossible possible worlds in giving semantics to the counterfactuals that arise in mathematical explanations. They also use causal models (which they call, as is also quite standard, *structural equations models*) to give semantics to counterfactuals. However, they do not use causality in their definition, so their analysis is quite different from the HP approach, which depends heavily on causality. Their focus is, roughly speaking, on how to construct the "impossible" possible worlds. They do not attempt to provide a way of determining whether one explanation is better than another.

Pincock (2023) criticizes causal approaches to mathematical explanation. Since he focuses on the definitions given by Lewis (2000) and Woodward (2003), some of the criticisms he makes do not apply to the HP approach. But his biggest concern is that these approaches do not give a particular notion of *mathematical* explanation (his principle four again) which, as I said, I am not trying to give.

Finally, it is worth stressing that the HP approach when combined with impossible possible worlds gives us a quantitative way of assessing when and why one explanation is better than another. To the best of my knowledge, while qualitative notions of what counts as a better explanation have been presented (see, e.g., (Woodward 2003)), no formal quantitative measures of "better explanation" have been proposed. It is clear that mathematicians do make such assessments. As Pincock [p. 34] (2023) says "mathematicians value some proofs because those proofs not only show that a theorem is true, but also explain why the theorem is true". While there has been interest in *explanatory proofs*, going back to Steiner (1978), and Pincock's fifth principle says "Some proofs of a theorem explain why that theorem is the case, while other proofs do not explain why that theorem is the case" (so that Pincock

does not want to identify proof with explanation), there seems to have been no attempt to quantify the extent to which a proof (or, more generally, an explanation) explains the statement that it is trying to explain. The definitions of explanatory power discussed in (Halpern 2016, Halpern and Pearl 2005b) apply to explanations of mathematical statements with no change. While the two simple examples that I focus on in this paper consider proofs of their conclusions from an empty set of premises, the same approach can be applied more generally to two different proofs from the same premises to the same conclusions. As I said, the equations in a causal model can be viewed as encoding the steps in the proof. A model can certainly be rich enough to encode the steps of several different proofs. We can then compare the explanatory power of these proofs.

The rest of this paper is organized as follows: in the next section, I briefly review causal models and the HP definition of explanation and explanatory power. In Section 3, I show how adding "impossible" causal models allows us to capture mathematical explanations.

§ 2. — Causal models and the HP definition of explanation.

In this section, I review causal models and the HP definition of explanation. The reader is encouraged to consult (Halpern 2016), from where this material is largely taken (almost verbatim), for further details. However, it should be possible to understand how I deal with mathematical explanations without understanding all the details of these definitions.

2.1. Causal models. Assume that the world is described in terms of variables and their values. Some variables may have a causal influence on others. This influence is modeled by a set of *structural equations*. It is conceptually useful to split the variables into two sets: the *exogenous* variables, whose values are determined by factors outside the model, and the *endogenous* variables, whose values are ultimately determined by the exogenous variables. The structural equations describe how these values are determined.

Formally, a *causal model* M is a pair $(\mathscr{S}, \mathscr{F})$, where $\mathscr{S}$ is a *signature*, which explicitly lists the endogenous and exogenous variables and characterizes their possible values, and $\mathscr{F}$ defines

a set of (*modifiable*) *structural equations*, relating the values of the variables. A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*). $\mathcal{F}$ associates with each endogenous variable $X \in \mathcal{V}$ a function denoted F_X (i.e., $F_X = \mathcal{F}(X)$) such that $F_X : (\times_{U \in \mathcal{U}} \mathcal{R}(U)) \times (\times_{Y \in \mathcal{V} - \{X\}} \mathcal{R}(Y)) \rightarrow \mathcal{R}(X)$. This mathematical notation just makes precise the fact that F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. I typically simplify notation and write $X = Y + U$ instead of $F_X(Y, Y', U) = Y + U$. (The fact that Y' does not appear on the right-hand side of the equation means that the value of X does not depend on Y'.)

The structural equations define what happens in the presence of external interventions. Setting the value of some set of variables $\vec{X}$ to $\vec{x}$ in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{\vec{X} \leftarrow \vec{x}}$, which is identical to M, except that the equations for variables in $\vec{X}$ in $\mathcal{F}$ are replaced by $X = x$ for each $X \in \vec{X}$ and its corresponding value $x \in \vec{x}$.

A variable Y *depends on* X if there is some setting of all the variables in $\mathcal{U} \cup \mathcal{V}$ other than X and Y such that varying the value of X in that setting results in a variation in the value of Y; that is, there is a setting $\vec{z}$ of the variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}) \neq F_Y(x', \vec{z})$. A causal model M is *recursive* (or *acyclic*) if there is no cycle of dependencies. It should be clear that if M is an acyclic causal model, then given a *context*, that is, a setting $\vec{u}$ for the exogenous variables in $\mathcal{U}$, the values of all the other variables are determined (i.e., there is a unique solution to all the equations). We can determine these values by starting at the top of the graph and working our way down. In this paper, following the literature, I restrict to recursive models. A pair $(M, \vec{u})$ consisting of a causal model M and a context $\vec{u}$ is called a (*causal*) *setting*.

A *causal formula* (*over* $\mathcal{S}$) is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where

- φ is a Boolean combination of *primitive events* (formulas of the form $X = x$),
- $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and
- $y_i \in \mathcal{R}(Y_i)$.

Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A causal formula ψ is true or false in a setting. As usual, I write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the setting $(M, \vec{u})$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with acyclic models) solution to the equations in M in context $\vec{u}$ (that is, the unique vector of values for the endogenous variables that simultaneously satisfies all equations in M with the variables in $\mathscr{U}$ set to $\vec{u}$). Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y}=\vec{y}}, \vec{u}) \models \varphi$.

It is worth noting that the choice of language (specifically, the set of endogenous and exogenous variables and their values) has a major impact on the set of possible explanations: a statement can't be an explanation if it cannot be expressed in the language.

2.2. Actual causality. A standard use of causal models is to define *actual causation*: that is, what it means for some particular event that occurred to cause another particular event. There have been a number of definitions of actual causation given for acyclic models (e.g., (Beckers 2021, Glymour and Wimberly 2007, Hall 2007, Halpern and Pearl 2005a, Halpern 2016, Hitchcock 2001, Hitchcock 2007, Weslake 2015, Woodward 2003)). Although most of what I say in the remainder of the paper applies without change to other definitions of actual causality in causal models, for definiteness, I focus here on what has been called the *modified* Halpern-Pearl definition (Halpern 2015, Halpern 2016), which I briefly review. (See (Halpern 2016) for more intuition and motivation.)

The events that can be causes are arbitrary conjunctions of primitive events; the events that can be caused are arbitrary Boolean combinations of primitive events. The definition takes as its point of departure the notion of *but-for* causality, widely used in the law; the intuition for but-for causality is that A is a cause of B if, had A not occurred, B would not have occurred. However, as is well known, the but-for test is not always sufficient to determine causality. Consider the following well-known example, taken from (Paul and Hall 2013):

> Suzy and Billy both pick up rocks and throw them at a bottle. Suzy's rock gets there first, shattering the bottle. Since both throws are perfectly accurate, Billy's would have shattered the bottle had it not been preempted by Suzy's throw.

Here the but-for test fails. Even if Suzy hadn't thrown, the bottle would have shattered. Nevertheless, I want to call Suzy's throw a cause of the bottle shattering. The following definition allows for causality beyond but-for causality.

Definition 2.1: $\vec{X} = \vec{x}$ is an *actual cause* of φ in $(M, \vec{u})$ if the following three conditions hold:

AC1. $(M, \vec{u}) \models (\vec{X} = \vec{x})$ and $(M, \vec{u}) \models \varphi$.

AC2. There is a set $\vec{W}$ of variables in $\mathcal{V}$ and a setting $\vec{x}'$ of the variables in $\vec{X}$ such that if $(M, \vec{u}) \models \vec{W} = \vec{w}$, then

$$(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}', \vec{W} \leftarrow \vec{w}]\neg\varphi.$$

AC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions AC1 and AC2. ∎

AC1 just says that $\vec{X} = \vec{x}$ cannot be considered a cause of φ unless both $\vec{X} = \vec{x}$ and φ actually happen. AC3 is a minimality condition, which says that a cause has no irrelevant conjuncts. AC2 captures the standard but-for condition ($\vec{X} = \vec{x}$ is a cause of φ if, had $\vec{X}$ been $\vec{x}'$ rather than $\vec{x}$, φ would not have happened) but allows us to apply it while keeping fixed some variables (i.e., the variables in $\vec{W}$) to the value that they had in the actual setting $(M, \vec{u})$. Note that if $\vec{W} = \varnothing$, then we get but-for causality. Thus, this definition generalizes but-for causality (although, for the examples in this paper, but-for causality suffices).

Example 2.2: Suppose that we want to model the fact that if an arsonist drops a match or lightning strikes then a forest fire starts. We could use endogenous binary variables *MD* (which is 1 if the arsonist drops a match, and 0 if he doesn't), *L* (which is 1 if lightning strikes, and 0 if it doesn't), and *FF* (which is 1 if there is a forest fire, and 0 otherwise), with the equation $FF = \max(L, MD)$; that is, the value of the variable *FF* is the maximum of the values of the variables *MD* and *L*. This equation says, among other things, that if $MD = 0$ and $L = 1$, then $FF = 1$. Alternatively, if we want to model the fact that a fire requires both a lightning strike *and* a dropped match (perhaps the wood is so wet that it needs two sources of fire to get going), then the only change in the model is that the equation for *FF* becomes $FF = \min(L, MD)$; the value of *FF* is the minimum of the values of *MD* and *L*. The only way that $FF = 1$ is if both $L = 1$ and $MD = 1$.

There is also an exogenous variable U that determines whether the lightning strikes and whether the match is dropped. U can take on four possible values of the form (i, j), where i and j are each either 0 or 1. Intuitively, i describes whether the external conditions are such that the lightning strikes (and encapsulates all such conditions, e.g., humidity and temperature), and j describes whether the arsonist drops the match (and thus encapsulates all the psychological conditions that determine this).

Consider the context where $U = (1, 1)$, so the arsonist drops a match and the lightning strikes. In the *conjunctive* model, where a fire requires both a lightning strike and a dropped match, it is easy to check that both $L =$ and $MD = 1$ are (separately) causes of $FF = 1$. Indeed, both $L = 1$ and $MD = 1$ are but-for causes. On the other hand, in the disjunctive model, where either $L = 1$ or $MD = 1$ suffices for the fire, neither $L = 1$ nor $MD = 1$ is a cause (since changing either one does not result in there not being a fire, no matter what we fix); the cause of the fire is the conjunction $L = 1 \wedge MD = 1$. As we shall see, things go the other way when it comes to explanation.

2.3. The HP definition of explanation. Unlike causality, as noted in the introduction, it is well known that what counts as an explanation depends on what the agent knows (Gärdenfors 1988, Salmon 1984). Here we model an agent's knowledge by means of an *epistemic state* $(\mathscr{K}, \Pr)$, where $\mathscr{K}$ is a set of causal settings and Pr is a probability on them. Intuitively, the causal settings in $\mathscr{K}$ are the ones that the agent considers possible, and reflects the agent's uncertainty regarding how the world works (represented by the equations in a model M in causal setting $(M, \vec{u})$) and what is currently true (represented by the context $\vec{u}$). As is standard, we say that an agent *knows* φ if φ is true at all the settings in $\mathscr{K}$. For simplicity, I assume that $\Pr(M, \vec{u}) > 0$ for each causal setting $(M, \vec{u}) \in \mathscr{K}$. The basic definition of explanation does not make use of Pr, just $\mathscr{K}$.[(1)]

I now give the formal definition, and then give intuition for the clauses, particularly EX1(a).

Definition 2.3: $\vec{X} = \vec{x}$ *is an explanation of* φ *relative to a set* $\mathscr{K}$ *of causal settings* if the following conditions hold:

[(1)] The definition given here is taken from (Halpern 2016), and differs slightly from the original definition given in (Halpern and Pearl 2005b).

EX1(a). If $(M, \vec{u}) \in \mathscr{K}$ and $(M, \vec{u}) \models \vec{X} = \vec{x} \wedge \varphi$, then there exists a conjunct $X = x$ of $\vec{X} = \vec{x}$ and a (possibly empty) conjunction $\vec{Y} = \vec{y}$ such that $X = x \wedge \vec{Y} = \vec{y}$ is a cause of φ in $(M, \vec{u})$.

EX1(b). $(M', \vec{u}') \models [\vec{X} \leftarrow \vec{x}]\varphi$ for all settings $(M', \vec{u}') \in \mathscr{K}$.

EX2. $\vec{X}$ is minimal; there is no strict subset $\vec{X}'$ of $\vec{X}$ such that $\vec{X}' = \vec{x}'$ satisfies EX1(a) and EX1(b), where $\vec{x}'$ is the restriction of $\vec{x}$ to the variables in $\vec{X}$.

EX3. For some $(M, \vec{u}) \in \mathscr{K}$, we have that $(M, \vec{u})) \models \vec{X} = \vec{x} \wedge \varphi$. (The agent considers possible a setting where the explanation and explanandum both hold.)

The explanation is *nontrivial* if it satisfies

EX4. $(M', \vec{u}') \models \neg(\vec{X} = \vec{x})$ for some $(M', \vec{u}') \in \mathscr{K}$ such that $(M', \vec{u}') \models \varphi$. (The explanation is not already known given the observation of φ.) ∎

The key part of the definition is EX1(b). Roughly speaking, it says that the explanation $\vec{X} = \vec{x}$ is a *sufficient cause* for φ: for all settings that the agent considers possible, intervening to set $\vec{X}$ to $\vec{x}$ results in φ. (See (Halpern 2016, Chapter 2.6) for a formal definition of sufficient cause.) EX2 and EX3 should be fairly clear.

EX4 is meant to capture the intuition (discussed in the introduction) that what counts as an explanation depends on what the agent knows. In the formal model, this is captured by the set $\mathscr{K}$. But what is $\mathscr{K}$ if we are looking for explanations of mathematical statements? If the primitive events are statements of mathematics (e.g., 4373 is prime), then if we insist that the possible worlds are all consistent with the facts of mathematics, there would be only one possible world and the agent would know all facts of mathematics. In this case, all explanations are known, and EX4 would not hold. But once we allow "impossible" possible worlds, $\mathscr{K}$ is no longer a singleton, even if we restrict the language to talking about mathematical statements. We allow the agent to consider it possible that 4373 is not prime. This makes it possible to satisfy EX4. Note that it means that whether "4373 is prime" is part of an explanation now depends on the agent's epistemic state. This seems to me consistent with the observation that when we explain a mathematical theorem to others, we must tailor the explanation to what they know (or understand).

That, of course, leaves open the question of what $\mathscr{K}$ should be. Whatever choice we make, we need to argue that it accurately represents the agent that we are trying to model. A poor choice would lead us to "explanations" that would not be useful to the agent to whom are trying to explain things. Baron, Colyvan, and Ripley (2017) and Kasirzadeh (2023) can be viewed as focusing on what the worlds in $\mathscr{K}$ should look like, and in particular, what mathematical facts should be true in an impossible possible world. Here I largely avoid that issue; by restricting the language to only those facts immediately relevant to the argument, I avoid the need to worry about what else is true. I return to this point at the end of Section 3.

That leaves EX1(a). Roughly speaking, it says that the explanation causes the explanandum. But there is a tension between EX1(a) and EX1(b) here: we may need to add conjuncts to the explanation to ensure that it suffices to make φ true in all contexts (as required by EX1(b)). But these extra conjuncts may not be necessary to get causality in all contexts. At least one of the conjuncts of $\vec{X} = \vec{x}$ must be part of a cause of φ, but the cause can include extra conjuncts. To understand why, it is perhaps best to look at an example.

Example 2.4: Going back to the forest-fire example, consider the following four contexts: in $u_0 = (0,0)$, there is no lightning and no arsonist; in $u_1 = (1,0)$, there is only lightning; in $u_2 = (0,1)$, the arsonist drops a match but there is no lightning; and in $u_3 = (1,1)$, there is lightning and the arsonist drops a match. Let M^d be the disjunctive model (where either lightning or a match suffices to start the forest fire), and let M^c be the conjunctive model (where we need both). Let $\mathscr{K}_1 = \{(M^d, u_0), (M^d, u_1), (M^d, u_2), (M^d, u_3)\}$. Both $L = 1$ and $MD = 1$ are explanations of $FF = 1$ relative to $\mathscr{K}_1$. Clearly EX1(b), EX2, EX3, and EX4 hold. For EX1(a), recall that in the setting (M^d, u_3), the actual cause is $L = 1 \wedge MD = 1$. Thus, EX1(a) is satisfied by $L = 1$ by taking $\vec{Y} = \vec{y}$ to be $MD = 1$, and is satisfied by $MD = 1$ by taking $\vec{Y} = y$ to be $L = 1$.

Now consider $\mathscr{K}_2 = \{(M^c, u_0), (M^c, u_1), (M^c, u_2), (M^c, u_3)\}$. The only explanation of fire relative to $\mathscr{K}_2$ is $L = 1 \wedge MD = 1$; due to the sufficiency requirement EX1(b), we need both conjuncts.

To take just one more example, if $\mathscr{K}_3 = \{(M^c, u_1), (M^c, u_3)\}$, then $MD = 1$ is an explanation of the forest fire. Since $L = 1$ is already known, $MD = 1$ is all the additional information that the agent needs to explain the fire. This is a trivial explanation: since

both $MD = 1$ and $L = 1$ are required for there to be a fire, the agent knows $MD = 1$ when he see the fire. Note that $MD = 1 \wedge L = 1$ is not an explanation; it violates the minimality condition EX2. $L = 1$ is not an explanation either, since $(M^c, u_1) \models \neg[L \leftarrow 1](FF = 1)$, so sufficient causality does not hold. ∎

2.4. Partial explanations and better explanations. Not all explanations are considered equally good. Moreover, it may be hard to find an explanation that satisfies EX1 for all settings $(M, \vec{u}) \in \mathscr{K}$; we may be satisfied with a formula that satisfies these conditions for almost all settings. In (Halpern 2016), various dimensions along which one explanation might be better than another are discussed. I focus on two of them here. Here the probability Pr in the epistemic state $(\mathscr{K}, \Pr)$ comes into play. Given a causal formula φ, let $[\![\varphi]\!]_{\mathscr{K}} = \{(M, \vec{u}) \in \mathscr{K} : (M, \vec{u}) \models \varphi\}$. If both $\vec{X} = \vec{x}$ and $\vec{X}' = \vec{x}'$ are explanations of φ relative to $\mathscr{K}$, $\vec{X} = \vec{x}$ is preferred relative to epistemic state $(\mathscr{K}, \Pr)$ if $\Pr([\![\vec{X} = \vec{x}]\!]_{\mathscr{K}} \mid [\![\varphi]\!]_{\mathscr{K}}) > \Pr([\![\vec{X}' = \vec{x}']\!]_{\mathscr{K}} \mid [\![\varphi]\!]_{\mathscr{K}})$, that is, if its conditional probability is higher. For example, if $\Pr(\{(M^d, u_1), (M^d, u_3)\}) > \Pr(\{(M^d, u_2), (M^d, u_3)\})$, then $L = 1$ would be viewed as a better explanation than $MD = 1$ along this dimension; it is more likely conditional on $[\![\varphi]\!]_{\mathscr{K}}$ (or, equivalently, likely, since $[\![\varphi]\!]_{\mathscr{K}}$ has probability 1).

Another consideration takes as its point of departure the fact that the conditions EX1(a) and (b) are rather stringent. We might consider $\vec{X} = \vec{x}$ quite a good explanation of φ relative to $\mathscr{K}$ if, with high probability, these conditions hold. More precisely, for EX1(a), consider the probability of the set of settings $(M, \vec{u})$ in $\mathscr{K}$ for which there exists a conjunct $X = x$ of $\vec{X} = \vec{x}$ and a (possibly empty) conjunction $\vec{Y} = \vec{y}$ such that $X = x \wedge \vec{Y} = \vec{y}$ is a cause of φ in $(M, \vec{u})$, conditional on $\vec{X} = \vec{x} \wedge \varphi$; for EX1(b), consider the probability of the set of settings $(M, \vec{u}) \in \mathscr{K}$ for which $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}]\varphi$. The higher these probabilities, the better the explanation. This also allows us to consider a *partial explanation*, one which satisfies EX1 with a probability less than 1.

§ 3. — Dealing with explanations of mathematical statements.

We can now return to the questions that motivated this paper. If we are looking for a (causal) explanation of "Why is 4373 the

sum of two squares?", we need to start with a causal model. If the explanation is going to be "it is a prime congruent to 1 mod 4", the model needs to include variables P_{4373}, $1M4_{4373}$, $S2S_{4373}$, where the first has value 1 if 4373 is an odd prime and 0 otherwise; the second has value 1 if 4373 is congruent to 1 mod 4, and 0 otherwise; and the third has value 1 if 4373 is the sum of two squares and 0 otherwise.

Now it is a fact of mathematics that 4373 is prime, congruent to 1 mod 4, and the sum of two squares, but since the agent does not necessarily know that (under some reasonable interpretation of the word "know"), the agent can consider models where at least P_{4373} and $S2S_{4373}$ take on value 0. (I am implicitly assuming that computing whether a number is congruent to 1 mod 4 is so simple that the agent does in fact know that $1M4_{4373} = 1$; nothing in the following discussion would change if the agent does not know this.) Assume that the value of P_{4373} and $1M4_{4373}$ is determined by an exogenous variable U that takes on four possible values of the form (i, j), where $i, j \in \{0, 1\}$, as in the forest-fire example; the value of i determines the value of P_{4373} and the value of j determines the value of $1M4_{4373}$.

If we further assume that the agent knows Fermat's two-squares theorem, then in all models that the agent considers possible, if $P_{4373} = 1$ and $1M4_{4373} = 1$, then $S2S_{4373} = 1$, and if $P_{4373} = 1$ and $1M4_{4373} = 0$, then $S2S_{4373} = 0$. The question is what should happen if $P_{4373} = 0$. While it is not hard to show that no number congruent to 3 mod 4 is the sum of two squares, there are non-primes congruent to 0, 1, and 2 mod 4, respectively, that are the sum of two squares (like 8, 25, and 18, respectively), and non-primes congruent to 0, 1, and 2 mod 4 that are not the sum of two squares (like 12, 33, and 6, respectively).

It seems reasonable to assume that the agent is uncertain about the effect of setting P_{4373} to 0 on $S2S$. (Recall that although this would result in an "impossible" possible world, the agent still considers it possible.) Thus, I consider two causal models that differ only in what happens if $P_{4373} = 0$: in one of them, call it M_1, it results in $S2S_{4373} = 0$; in the other, call it M_2, results in $S2S_{4373} = 1$. (Otherwise, the models are identical.) Let u_0, u_1, u_2, and u_3 be the contexts where U takes on values $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$, respectively. Since I am assuming that the agent knows that $1M4_{4373} = 1$, I take $\mathscr{K} = \{(M_1, u_1), (M_1, u_3), (M_2, u_1), (M_2, u_3)\}$.

It is now straightforward to show that, relative to $\mathscr{K}$, the fact that $P_{4373} = 1 \wedge 1M4_{4373} = 1$ (i.e., the fact the 4373 is an odd prime and congruent to 1 mod 4) is an explanation of the fact that 4373

is the sum of two squares. Let us go through the conditions in Definition 2.3. For EX1(a), there are only two settings in $\mathscr{K}$ that satisfy $P_{4373} = 1 \wedge 1M4_{4373} = 1$, namely (M_1, u_3) and (M_2, u_3). In both of these settings, $1M4_{4373} = 1$ is a cause of $S2S_{4373} = 1$. It is in fact a but-for cause: if we set $1M4_{4373} = 0$, then $S2S_{4373} = 0$ (since $P_{4373} = 1$ continues to hold). EX1(b) is immediate, since if 4373 is an odd prime that is congruent to 1 mod 4, then it is guaranteed to be the sum of two squares (the equations in M_1 and M_2 enforce this). For EX2, note that $P_{4373} = 1$ does not satisfy EX1(a) and $1M4_{4373} = 1$ does not satisfy EX1(b). EX3 clearly holds, since $P_{4373} = 1 \wedge 1M4_{4373} = 1 \wedge S2S_{4373}$ holds in both (M_1, u_3) and (M_2, u_3). Finally, EX4 also holds; $\neg(P_{4373} = 1 \wedge 1M4_{4373} = 1)$ holds in (M_1, u_1) and (M_2, u_1).

Although $1M4_{4373} = 1$ is known (it is true in all settings in $\mathscr{K}$), it is part of the explanation. In general, an explanation will not include facts that an agent already knows unless these facts are necessary to show causality (i.e., EX1(a)). Nothing would have changed if $1M4_{4373} = 1$ were not known (i.e., if (M_1, u_0), (M_1, u_2), (M_2, u_0), and (M_2, u_2) were in $\mathscr{K}$): $P_{4373} = 1 \wedge 1M4_{4373} = 1$ would still have been an explanation of $S2S_{4373}$.

Now consider the other possible explanation of 4373 being the sum of two squares, namely the demonstration that $4373 = 62^2 + 23^2$. We could capture this by adding another variable to the model, $S2S_{4373,62,23}$, such that $S2S_{4373,62,23} = 1$ if $4373 = 62^2 + 23^2$ and 0 otherwise. Again, the agent can consider it possible that $S2S_{4373,62,23} = 0$. Setting $S2S_{4373,62,23} = 1$ should result in $S2S_{4373} = 1$, independent of the values of the other variables (although, in light of the theorem, the agent would not consider possible a model where $S2S_{4373,62,23} = 1$, $P_{4373} = 1$, and $1M4_{4373} = 0$). But what if $S2S_{4373,62,23} = 0$? Assume that in this case, the agent is in the same situation he was in before, and considers all the settings in $\mathscr{K}$ possible. More precisely, let M_1' and M_2' be like M_1 and M_2, respectively, except that (1) the exogenous variable U' now has values of the form (i, j, k), for $i, j, k \in \{0, 1\}$, and determines P_{4373}, $1M4_{4373}$, and $S2S_{4373,62,23}$ in the obvious way; and (2) the equation for $S2S_{4373}$ is such that if $S2S_{4373,62,23} = 1$, then $S2S_{4373} = 1$ in both models, while if $S2S_{4373,62,23} = 0$, then the value of $S2S_{4373}$ is determined in M_1' (resp., M_2') in the same way that it is determined in M_1 (resp., M_2).

Let $u_0', \ldots, u_7'$ be the contexts where U' takes on values $(0,0,0)$, $(0,1,0)$, $(1,0,0)$, $(1,1,0)$, $(0,0,1)$, $(0,1,1)$, $(1,0,1)$, and $(1,1,1)$,

respectively; let $\mathscr{K}' = \{(M', u') : M' \in \{M'_1, M'_2\}, u' \in \{u'_1, u'_3, u'_5, u'_7\}\}$. $S2S_{4373,62,23} = 1$ is not an explanation of $S2S_{4373}$ relative to $\mathscr{K}'$ because it fails EX1(a). For example, $S2S_{4373,62,23} = 1$ is not a cause of $S2S_{4373} = 1$ in (M'_2, u'_7). Of course, this conclusion depends on how we modeled the set $\mathscr{K}'$ of possible worlds. As I noted earlier, the "right" choice of $\mathscr{K}'$ depends on the agent we are trying to model. While I think that the $\mathscr{K}'$ defined here is reasonable, in the sense that it is a reasonable representation of the beliefs of a typical agent, there are surely agents for whom other choices of $\mathscr{K}'$ would be appropriate; for such agents, $S2S_{4373,62,23} = 1$ might well be an explanation.

The analysis of the second example is similar. The fact that $x - 2$ is a factor of $f(x)$ is an explanation of the fact that $f(2) = 0$ (in the model where there is a variable F_{x-2} that is 1 if $x - 2$ ia a factor of $f(x)$ and 0 otherwise and a variable $f2E0$ that is 1 if $f(2) = 0$ and 0 otherwise). It seems reasonable to expect the agent to understand that $x - 2$ is factor of $f(x)$ iff $f(2) = 0$, and for the equation for $f2E0$ to reflect this, so the fact that $x - 2$ is a factor of f is a cause of $f2E0 = 1$. The situation for (1) is a little more subtle. It is not immediately transparent that the left-hand side of (1) is the result of plugging in 2 for x in the polynomial f. Thus, the agent might consider it possible that (1) not hold and yet $f(2) = 0$; more precisely, the agent would consider possible a model where (1) does not hold and $f(2) = 0$. In this model, (1) is not a cause of $f(2) = 0$, so if the agent's set of possible settings includes this model, (1) would not be an explanation of $f(2) = 0$. I would argue that at least part of the reason that people find the fact that $x - 2$ is a factor of f a better explanation for $f(2) = 0$ than (1) is because the fact that $f(2) = 0$ is not immediately obvious from (1).

Now consider the quality of the explanation. Although, as I have argued, the fact that $4373 = 62^2 + 23^2$, that is, $S2S_{4373,62,23} = 1$, is not an explanation of 4373 being the sum of two squares, it could be a good partial explanation. It does satisfy EX1(b) in all settings in $\mathscr{K}$, but satisfies EX1(a) only in (M'_1, u'_5), so how good a partial explanation it is depends on the probability of (M'_1, u'_5). By way of contrast, $P_{4373} = 1 \wedge 1M4_{43u73} = 1$ satisfies EX1(a) and EX1(b) in all settings, so no matter what Pr is, it is at least as good an explanation as $S2S_{4373,62,23} = 1$, and a strictly better explanation, at least as far as this consideration goes, if $\Pr(M'_1, u'_5) < 1$.

On the other hand, the probability of $P_{4373} = 1 \wedge 1M4_{43u73} = 1$ conditional on $S2S_{4373} = 1$ is not necessarily higher than the probability of $S2S_{4373,62,23} = 1$ conditional on $S2S_{4373} = 1$; it depends on the relative probability of $\{(M'_1, u'_3), (M'_1, u'_7), (M'_2, u'_3), (M'_2, u'_7)\}$ and $\{(M'_2, u'_5), M'_2, u'_7)\}$. While I have placed no constraints on Pr, there is at least a heuristic argument that the agent should consider the former event more likely than the latter: by the prime number theorem (Jameson 2003), there are roughly $\ln(n)$ prime numbers less than n; it seems reasonable to expect roughly half of them to be 1 mod 4. On the other hand, there is at most one number less than n that is $62^2 + 23^2$. If Pr respects this reasoning, then again, $P_{4373} = 1 \wedge 1M4_{4373} = 1$ is a better explanation of $S2S_{4373} = 1$ than $S2S_{4373,62,23} = 1$.

In the second example, considering EX1(a) and EX1(b), arguing just as in the first example, the fact that $x - 2$ is a factor of f is clearly at least as good an explanation as (1), and a strictly better explanation if the agent places positive probability on a setting where (1) does not hold and $f(2) = 0$. But now considering the probability conditional on $f(2) = 0$, the fact that $x - 2$ is a factor of f has probability 1 (since the agent is assumed to know that $f(2) = 0$ iff $x - 2$ is a factor of f), so again it is at least as good an explanation when considering the conditional probability, and a strictly better explanation if the agent places positive probability on a setting where (1) does not hold and $f(2) = 0$.

I close by considering an issue raised by Lange (2022). He points out that, with a counterfactual account, a given mathematical fact may have arguably too many explanations. He gives as an example the problem of explaining the fact that 123284 is divisible by 37. He points out that, according to the counterfactual account that he discusses, all of the following are explanations:

- $123321 = 123284 + 37$, 123284 is divisible by 37, and $a + 37$ is divisible by 37 iff a is divisible by 37.
- $123321 = 123358 - 37$, 123358 is divisible by 37, and $a - 37$ is divisible by 37 iff a is divisible by 37.
- 444 is divisible by 37, 123 + 321 = 444, and a is divisible by 37 iff the number obtained by taking a's digits (in base 10) in groups of 3, beginning from the right, and adding those numbers together, is divisible by 37.

Each of these statements could potentially be an explanation in the framework presented here as well *provided that the relevant primitive events are in the language*.[(2)] We can think of the language as describing concepts that are salient to the agent. It seems to me unlikely that the relevant facts in the explanations above are salient to the agent, but it is certainly not impossible. For example, suppose that we have, for some reason, been discussing the number 123284 and the fact that it is divisible by 37. In that case, it seems reasonable to take the second explanation as acceptable.

The approach taken here also gives us another way to deal with the plethora of possible explanation: not all explanations are equally good. While I have no formal proof of this fact, I believe that the goodness of explanations will align well with what working mathematicians take to be good explanations.

To summarize, what counts as an explanation under the approach proposed here is heavily dependent on the choice of language and what the agent knows. The goodness of an explanation (which gives us a way of preferring one explanation to another) also depends on the agent's epistemic state, and how likely the agent takes various impossible possible worlds to be. This seems to me completely consistent with how actual explanations work. When giving an explanation, we are (or should be) sensitive to what the agent to whom we are giving the explanation already knows and considers possible, and to using appropriate concepts.

§ — Bibliography.

Baron, S., M. Colyvan, and **D. Ripley** (2017). How mathematics can make a difference. *Philosophers' Imprint 17*(3), 1–19.

Beckers, S. (2021). Causal sufficiency and actual causation. *Journal of Philosophical Logic 50*, 1341–1374.

Cresswell, M. J. (1970). Classical intensional logics. *Theoria 36*, 347–372.

Cresswell, M. J. (1972). Intensional logics and logical truth. *Journal of Philosophical Logic 1*, 2–15.

[(2)] We would also have to remove statements that are already known; for example, if it is known that a is divisible by 37 iff $a + 37$ is divisible by 37, this would not be part of the explanation.

Cresswell, M. J. (1973). *Logics and Languages*. London: Methuen and Co.

Gärdenfors, P. (1988). *Knowledge in Flux*. Cambridge, MA: MIT Press.

Glymour, C. and **F. Wimberly** (2007). Actual causes and thought experiments. In J. Campbell, M. O'Rourke, and H. Silverstein (Eds.), *Causation and Explanation*, pp. 43–67. Cambridge, MA: MIT Press.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies 132*, 109–136.

Halpern, J. Y. (2015). A modification of the Halpern-Pearl definition of causality. In *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pp. 3022–3033.

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

Halpern, J. Y. and **J. Pearl** (2005a). Causes and explanations: a structural-model approach. Part I: causes. *British Journal for Philosophy of Science 56*(4), 843–887.

Halpern, J. Y. and **J. Pearl** (2005b). Causes and explanations: a structural-model approach. Part II: explanations. *British Journal for Philosophy of Science 56*(4), 889–911.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. New York: Free Press.

Hintikka, J. (1962). *Knowledge and Belief*. Ithaca, N.Y.: Cornell University Press.

Hintikka, J. (1975). Impossible possible worlds vindicated. *Journal of Philosophical Logic 4*, 475–484.

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

Hitchcock, C. (2007). Prevention, preemption, and the principle of sufficient reason. *Philosophical Review 116*, 495–532.

Jameson, G. J. O. (2003). *The Prime Number Theorem*. Cambridge University Press.

Kasirzadeh, A. (2023). Counter countermathematical explanations. *Erkenntnis 88*, 2537–2560.

Kripke, S. (1965). A semantical analysis of modal logic II: non-normal propositional calculi. In L. Henkin and A. Tarski (Eds.), *The Theory of Models*, pp. 206–220. Amsterdam: North-Holland.

Lange, M. (2022). Challenges facing counterfactual accounts of explanation in mathematics. *Philosophia Mathematica 30*(1), 32–58.

Lewis, D. (2000). Causation as influence. *Journal of Philosophy XCVII*(4), 182–197.

Paul, L. A. and **N. Hall** (2013). *Causation: A User's Guide*. Oxford, U.K.: Oxford University Press.

Pincock, C. (2023). *Mathematics and Explanation*. Cambridge University Press.

Rantala, V. (1982). Impossible worlds semantics and logical omniscience. *Acta Philosophica Fennica 35*, 18–24.

Salmon, W. C. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, N.J.: Princeton University Press.

Steiner, M. (1978). Mathematical explanation. *Philosophical Studies 34*, 135–151.

Weslake, B. (2015). A partial theory of actual causation. *British Journal for the Philosophy of Science*. To appear.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford, U.K.: Oxford University Press.

Joseph Y. Halpern
Computer Science Department
Cornell University
halpern@cs.cornell.edu