# Actual Causality: A Primer

Sankaran Vaidyanathan
sankaranv@cs.umass.edu

## 1 Introduction

A growing body of work in causal inference focuses on the question of inferring *actual causes* [Pearl, 2000, Halpern, 2016], which are causes of particular events in a specific observed context. Philosophers have typically distinguished two notions of causality: *type causality*, sometimes called general causality, makes general statements about causal relationships between entities in the world, but *actual causality*, sometimes called token causality or specific causality, makes statements about particular events. For example, type causality is used to make conclusions such as "smoking causes cancer", and actual causality is used to make conclusions such as "the fact that this person smoked one pack of cigarettes a day for 30 years caused him to develop the tumor".

The vast majority of work in causal inference has dealt with type causation, and its utility mainly lies in tasks that require making predictions. Knowledge of general causal relationships between entities in the world (or in a model) allow one to make inferences such as "if you smoke one pack of cigarettes a day, you are likely to get lung cancer". Prediction is the focus of many efforts in machine learning, and type causation is often used in applications where we are looking forward.

In contrast, actual causality is mostly concerned with applications that are looking backwards, and identifying causes that have already occurred for events that have already happened. In particular, actual causality is a critical component of blame and responsibility assignment, explanation, and safety assessment. A statement like "the accident was caused by the faulty brakes in the car, and not the pouring rain or the blood alcohol level of the driver" falls within the domain of actual causality. The intuitions behind many previous works in actual causality come from examples in law, where we often know the relevant facts but still need to establish causality. Both type and actual causality have important applications and are intertwined, but the latter has been studied far less in the various communities concerned with empirical research in causal inference.

Reasoning about actual causation is a principled and promising direction for research in AI, particularly for explanation, safety, and robustness. At any given instant there are likely to be many events occurring in the environment that are distracting and irrelevant to the current state of an agent. Disregarding these can aid in training more robust agents and generating more precise explanations or safety assessments, but type causality does not provide the capability to prune out these events on a case-by-case basis. Additionally, it is possible to make counterintuitive and incorrect inferences if applying type causal statements to actual causal queries, as we will demonstrate later in this paper. In previous work, actual causality has been used to formalize problems in AI such as explanation [Beckers, 2022], blame and responsibility [Chockler and Halpern, 2004], and assessment of harm [Richens et al., 2022, Beckers et al., 2022]. However, not much progress in this field has been made outside of the philosophy literature, and the ability to accurately and tractably infer actual causation is essential for it to be applied to real AI systems.

## 2 Where type and actual causation disagree

In general, statements about type causation cannot be directly applied to making inferences about actual causation. In particular, simple counterfactual reasoning with a causal model that was trained to identify type causes can give counterintuitive results when applied to actual causation. This is mostly illustrated in the philosophy literature using examples, such as throwing rocks at a bottle or a gang of shooters firing at a single target, but the implications of these examples are more generally applicable to scenarios in ML and RL environments.

The most common problems where type causal inferences fail to correctly establish actual causality are *preemption* problems, *overdetermination* problems, and *normality* problems. Preemption occurs when there are multiple causes of an outcome but one always precedes the other. In these cases, the type causal relationship between the slower cause and the outcome will appear or disappear based on the specific assignment to the faster cause. Rock throwing is a classic example of preemption. Overdetermination occurs when there are multiple simultaneous causes for the same outcome, and it is not clear whether any one of them is necessary. For example, if there are three different events and any one of them was sufficient to cause the outcome, we would not see a causal effect between any single event and the outcome because the other two would always guarantee that the outcome is unchanged. Forest fire and gang shooting are classic examples of this. Normality problems occur when there are multiple candidate causes but one of them is clearly less plausible than the other. The Queen of England problem and the agent and obstacle problem are examples of this. All of these examples will be discussed in the sections that follow.

One reason for the appearance of these examples is because while many variables can have general causal relationships with a given outcome variable, in a specific observed context only a select few might actually be responsible for the outcome that was observed. Depending on the nature of the model or world, some type causal relationships actually appear only in certain instances. In general, the problem of inferring causation in the presence of context-specific independences is inextricably tied to actual causality. For example, if an agent is trying to push a block and there are obstacles present in the environment, all the obstacles have causal relationships with the block being pushed but they only become active when they are at a certain level of proximity to the block.

## 3 Notation for deterministic SCMs

The question of whether an event is an actual cause of some outcome or not is usually determined with respect to a model $M$ and a specific setting of the exogenous variables $\mathbf{u}$; the tuple $(M, \mathbf{u})$ is jointly referred to as a *causal setting*, and $\mathbf{u}$ is referred to as a *context*. Halpern [2016] emphasizes that causation is defined with respect to a model, and hence we must assume that all relevant facts about the world in which we are trying to reason are given in the model and context $(M, \mathbf{u})$.

A *structural causal model* is a tuple $M = (\mathcal{U}, \mathcal{V}, \mathcal{F})$, where $\mathcal{U}$ is a set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{F}$ is a set of structural equations, one for each $V$ in $\mathcal{V}$. Many papers additionally refer to a *signature* $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$ where $\mathcal{R}$ denotes the ranges of each of the variables in $\mathcal{U}$ and $\mathcal{V}$. For example, a binary variable $X$ will have $\mathcal{R}(X) = \{0, 1\}$.

The notation $(M, \mathbf{u}) \models X = x$ denotes that the variable $X$ will take on value $x$ once the exogenous variables are set to $\mathbf{u}$. This implicitly assumes that the model $M$ is *recursive*, or in other words its causal graph is *acyclic*. Therefore, the value of any endogenous variable $X$ can be obtained just using the context $\mathbf{u}$, by recursively applying each function in the SCM. Alternatively, we can express the value of any variable $X$ as a function of the exogenous variables alone, denoted $X(\mathbf{u})$.[1]

A *primitive event* is an assignment of a single endogenous variable $X = x$, and the outcome of interest $\varphi$ is a Boolean combination of primitive events. For a given set of endogenous variables $\mathbf{X} = \{X_1, ...X_k\}$ where $X_1, ...X_k \in \mathcal{V}$, and corresponding assignments $\mathbf{x} = \{x_1, ...x_k\}$ where $x_i \in \mathcal{R}(X_i)$ for all $i = 1, ..., k$, we can define a *causal formula* $\psi = [\mathbf{X} \leftarrow \mathbf{x}]\varphi$ to denote that $\varphi$ holds after an intervention that sets the value of $\mathbf{X}$ to $\mathbf{x}$. The arrow notation in $\mathbf{X} \leftarrow \mathbf{x}$ is meant to indicate an intervention, as opposed to the notation $\mathbf{X} = \mathbf{x}$ which only indicates that $\mathbf{x}$ is the observed value of $\mathbf{X}$.

Similarly, we can say that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}]\varphi$ if the outcome of interest $\varphi$ is true in the model that results from the intervention $\mathbf{X} \leftarrow \mathbf{x}$ in the same context $\mathbf{u}$. In other words, $\varphi$ holds when the exogenous variables are set to $\mathbf{u}$ and endogenous variables $\mathbf{X}$ are set to $\mathbf{x}$. The model that results from the intervention $\mathbf{X} \leftarrow \mathbf{x}$ is often written as $M_{\mathbf{X} \leftarrow \mathbf{x}}$, and we can therefore write

$$(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}]\varphi \text{ if and only if } (M_{\mathbf{X} \leftarrow \mathbf{x}}, \mathbf{u}) \models \varphi$$

Halpern [2016] mentions that the notation $(M, \mathbf{u}) \models \varphi$ is standard in the logic and philosophy literature, but differs from the more compact notation used in statistics and related communities. For example, $(M, \mathbf{u}) \models X = x$ may be written as $X(\mathbf{u}) = x$ or just $X = x$, and $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}](Y = y)$ may be written as $Y_x(\mathbf{u}) = y$. Halpern argues that these notations sacrifice clarity by suppressing the model $M$, assume only one outcome variable $Y$ and only one relevant variable $X$ to be intervened on, and if vector notation is used then the order of variables needs to be separately established.

Note: for the rest of this document, I will slightly depart from the notation used in the actual causality literature by writing the sets $\mathcal{U}$ and $\mathcal{V}$ instead as vectors $\mathbf{U}$ and $\mathbf{V}$, and generally use boldface to denote that the quantity is a tensor, so that they are easier to plug into equations from statistics, linear algebra, etc. that are relevant to machine learning.

# 4 Deterministic Actual Cause

**Definition 1** (General Definition of Causation)**.** *For a given model $M$, the event $\mathbf{X} = \mathbf{x}$ is an actual cause of the outcome $\mathbf{Y} = \mathbf{y}$ for a given context $\mathbf{U} = \mathbf{u}$ if there exists a witness set $\mathbf{W}$ with values $\mathbf{w}^*$ that satisfy the following conditions:*

- **AC1** (*Factual*): The given event $\mathbf{X} = \mathbf{x}$ and outcome $\mathbf{Y} = \mathbf{y}$ actually happened, or in other words $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})$.
- **AC2(a)** (*Necessity*): There is some contrast value $\mathbf{X} = \mathbf{x}'$ such that $(\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}^*)$ is not sufficient for $\mathbf{Y} = \mathbf{y}$.
- **AC2(b)** (*Sufficiency*): $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}^*)$ is sufficient for $\mathbf{Y} = \mathbf{y}$.

[1] This will be useful when extending to probabilistic SCMs, since we need to push the uncertainty forward from a distribution over the exogenous variables to obtain distributions over all the endogenous variables.

- **AC3** (*Minimality*): There is no strict subset $\mathbf{X}' \subset \mathbf{X}$ for which an assignment $\mathbf{X}' = \mathbf{x}'$ exists that satisfies both AC2(a) and AC2(b).

In the above definition, the assignment $\mathbf{w}^*$ indicates that the witness set can only take on its actual values, and AC2(a) formalizes the *contrastive necessity* definition from the philosophy literature. The goal of this definition is to capture the idea that for the event $\mathbf{X} = \mathbf{x}$ to cause the outcome $\mathbf{Y} = \mathbf{y}$, it must be a *necessary element of a sufficient set (NESS)* for the given outcome. However, this definition does not expand on what it means for an event to be sufficient for an outcome. The discussion that follows will cover different interpretations of sufficiency, the meaning of the witness set, and implications of these definitions.

# 5 Deterministic Causal Sufficiency

Informally, to say that some event $\mathbf{X} = \mathbf{x}$ is sufficient for the outcome $\mathbf{Y} = \mathbf{y}$ is to say that the latter follows from the former, and Pearl argues that existing logical notions of necessity and sufficiency lack the resources to formalize this. One obvious demand is that the meaning of causal sufficiency should capture the directionality of causation, which comes down to treating $\mathbf{X} = \mathbf{x}$ as an intervention and $\mathbf{Y} = \mathbf{y}$ as the consequence of that intervention.

However, this says nothing about the other endogenous variables $\mathbf{V} - (\mathbf{X} \cup \mathbf{Y})$ and their values, nor about the contexts $\mathbf{u}$ in which we are evaluating the intervention. The difficulty lies in deciding what conditions we choose to impose on the other variables, both endogenous and exogenous, and the conditions we choose can result in totally different definitions and consequences. We will review a few different candidates for these conditions, based on the definitions from Beckers [2021].

## 5.1 Direct Sufficiency

$\mathbf{X} = \mathbf{x}$ is *directly sufficient* for $\mathbf{Y} = \mathbf{y}$ in $M$ if for all $\mathbf{c} \in \mathcal{R}(\mathbf{C})$ and all $\mathbf{u} \in \mathcal{R}(\mathbf{U})$ we have that $(M, \mathbf{u}) \models [\mathbf{X} \leftarrow \mathbf{x}, \mathbf{C} \leftarrow \mathbf{c}]\mathbf{Y} = \mathbf{y}$.

This is the strongest possible condition: in every possible context $\mathbf{u}$, if we apply the intervention $\mathbf{X} \leftarrow \mathbf{x}$ then the outcome will be $\mathbf{Y} = \mathbf{y}$, irrespective of the values of all other variables. In other words, the outcome is robust to all possible interventions on the remaining variables $\mathbf{C} = \mathbf{V} - (\mathbf{X} \cup \mathbf{W} \cup \mathbf{Y})$.

If we are given a single context $\mathbf{u}$, we can remove the requirement that the given condition should hold for all $\mathbf{u} \in \mathcal{R}(\mathbf{U})$. In Beckers [2021], the definition for a given assignment of $\mathbf{U}$ is called *actual direct sufficiency*, and the condition where the assignment of $\mathbf{U}$ is unknown is denoted as simply *direct sufficiency*. For simplicity, we will just write the definitions for unknown $\mathbf{U}$ going forward, since this is more relevant for the probabilistic definitions we aim to develop, and the actual versions of the definitions are implied by simply dropping the for-all quantifier over $\mathbf{U}$.

We can show that once minimality is enforced, only parents of $\mathbf{Y}$ can be part of the set $\mathbf{X}$ under this definition. To see this, say that $Y = A$, $A = X$ and $X = U$, and the context says that $U = 1$. $X = 1$ is not directly sufficient for $Y = 1$ because intervening on $A$ would override the influence of $X$ on $Y$.

### 5.2 Weak Sufficiency

$\mathbf{X} = \mathbf{x}$ is *weakly sufficient* for $\mathbf{Y} = \mathbf{y}$ in $M$ if for all $\mathbf{u} \in \mathcal{R}(\mathbf{U})$ we have $(M, \mathbf{u}) \vDash [\mathbf{X} \leftarrow \mathbf{x}]\mathbf{Y} = \mathbf{y}$.

This is a much weaker condition: in every possible context $\mathbf{u}$, when the intervention $\mathbf{X} \leftarrow \mathbf{x}$ is applied and *no other intervention is applied*, we will always get $\mathbf{Y} = \mathbf{y}$. This assumes that the other variables take on their observational values, and we can solve for them by substituting $\mathbf{u}, \mathbf{x}$ and $\mathbf{y}$ into the equations.

We can see that under this definition and only one context $\mathbf{u}$, the sufficiency condition AC2(b) is trivially satisfied if AC1 is satisfied. This is because for a given $\mathbf{u}$, if you replace the equations for $\mathbf{W}$ and $\mathbf{X}$ with their actual values then all the other variables in $\mathbf{V}$ would still take on their actual values. The remaining three conditions make up the Modified HP definition from Halpern [2016], which is the most widely known definition of actual cause.

**Definition 2** (Modified HP)**.** *For a given model $M$, the event $\mathbf{X} = \mathbf{x}$ is an actual cause of the outcome $\mathbf{Y} = \mathbf{y}$ for a given context $\mathbf{U} = \mathbf{u}$ if:*

- **AC1** (*Factual*): The given event $\mathbf{X} = \mathbf{x}$ and outcome $\mathbf{Y} = \mathbf{y}$ actually happened, or in other words $(M, \mathbf{u}) \vDash (\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y})$.
- **AC2** (*Necessity*): There exists a witness set $\mathbf{W}$ such that when they are fixed to their actual values $\mathbf{w}^*$, there exists a contrast value $\mathbf{X} = \mathbf{x}'$ such that $(\mathbf{X} = \mathbf{x}', \mathbf{W} = \mathbf{w}^*)$ is not weakly sufficient for $\mathbf{Y} = \mathbf{y}$. In other words we will have $(M, \mathbf{u}) \vDash [\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}](\mathbf{Y} \neq \mathbf{y})$, assuming all other variables take on their observational values.
- **AC3** (*Minimality*): There is no strict subset $\mathbf{X}' \subset \mathbf{X}$ for which an assignment $\mathbf{X}' = \mathbf{x}'$ exists that satisfies AC2.

The weak sufficiency definition, while allowing for actual causes that are not direct parents, can have counterintuitive implications. For example, if two variables $X$ and $Y$ have no path to each other, like say $X = U$ and $Y = 1$, any value of $X$ will be weakly sufficient for $Y = 1$ because nothing can change the outcome. This type of example has implications for RL and simulation environments, for example in a setting where there is an agent pushing a block and an obstacle is present in the map. According to the Modified HP definition, the obstacle will always be an actual cause of the block's movement even if the obstacle is somewhere far away. This is because by including the agent in the witness set, the obstacle's motion is weakly sufficient because the agent pushed the block and it actually moved. The obstacle's motion is also necessary, because the counterfactual setting of the obstacle position where it is put in front of the block will make it not possible to move. Intuitively, we would not want the obstacle to be an actual cause since we know it is unrelated to the block's motion.

We can also contrast the above two definitions using the following example, based on Karimi et al. [2021] and Beckers [2022]. Consider a system for loan applications that is captured by a causal model such that $Y = (X_1 + 5X_2 - 225{,}000) > 0$, where $Y$ is a binary variable representing whether the loan is granted, $X_1$ is the applicant's income, and $X_2$ is the applicant's savings. Further, assume that the applicant's savings are determined by their initial deposit $X_3$ and their income in the following manner: $X_2 = 3/10X_1 + X_3$. It is also the case that people with high savings take out a safety deposit box ($X_4$) at the bank: $X_4 = (X_2 > 1{,}000{,}000)$.

It can be verified that $X_2 = 45,001$ is both weakly and directly sufficient for $Y = 1$. However, $(X_1 = 50,000, X_3 = 25,000)$ is weakly sufficient but not directly sufficient for $Y = 1$, and it is important to know which definition to use since their implications are different. For example, an applicant for a loan who is told that their income $(X_1)$ and initial deposit $(X_3)$ are sufficient (based on the weak sufficiency definition) may incorrectly conclude that they can spend $20,000$, and doing so will get their loan application denied.

## 5.3 Strong Sufficiency

This definition, introduced in Beckers [2021] but based on a definition from Weslake [2015], is meant to address the need for a definition that lies in between weak and strong sufficiency. The proposed definition is the transitive closure of direct sufficiency, written as follows:

$\mathbf{X} = \mathbf{x}$ is *strongly sufficient* for $\mathbf{Y} = \mathbf{y}$ in $M$ along a *network* $\mathbf{N}$ if there exists a set $\mathbf{N}$ that contains $\mathbf{Y}$ (i.e. $\mathbf{Y} \subseteq \mathbf{N}$) and a corresponding assignment $\mathbf{N} = \mathbf{n}$ which contains $\mathbf{Y} = \mathbf{y}$, such that $\mathbf{X} = \mathbf{x}$ is directly sufficient for $\mathbf{N} = \mathbf{n}$. In other words, $\mathbf{Y} = \mathbf{y}$ is part of a set for which $\mathbf{X} = \mathbf{x}$ is directly sufficient, even if this sufficiency condition does not hold for $\mathbf{Y} = \mathbf{y}$ alone. In the context of generating explanations, the variables in the network $\mathbf{N}$ are assumed to be safeguarded from interventions, while the outcome is expected to be robust to all possible interventions on the remaining variables.

An equivalent definition is that there are (possibly overlapping) sets $\mathbf{N_i}$ such that $\mathbf{N} = \mathbf{N_1} \cup ... \cup \mathbf{N_k} \cup \mathbf{Y}$ and there exist values $\mathbf{n_i} \in \mathcal{R}(\mathbf{N_i})$ for each $i \in \{1, ..., k\}$ such that $\mathbf{X} = \mathbf{x}$ is directly sufficient for $\mathbf{N_1} = \mathbf{n_1}$, $\mathbf{N_1} = \mathbf{n_1}$ is directly sufficient for $\mathbf{N_2} = \mathbf{n_2}$,..., and $\mathbf{N_k} = \mathbf{n_k}$ is directly sufficient for $\mathbf{Y} = \mathbf{y}$. The network can be thought of as variables that lie on a path from $\mathbf{X}$ to $\mathbf{Y}$.

To see this, we return to our previous example where $Y = A$, $A = X$ and $X = U$, and the context says that $U = 1$. $X = 1$ was not directly sufficient for $Y = 1$ because intervening on $A$ would override the influence of $X$ on $Y$, but we can say that $X = 1$ is directly sufficient for $(A = 1, Y = 1)$, or equivalently $X = 1$ is strongly sufficient for $Y = 1$ through the network $A = 1$. Because the set of other endogenous variables $\mathbf{C}$ is empty and we do not have to consider any values of $(A, Y, X)$, we can see that when we intervene and set $X = 1$ we get $Y = A = X = 1$. The premise of the strong sufficiency definition is that the fact that $X = 1$ is directly sufficient for some assignment of a set of variables that includes $Y$ is still a meaningful causal relationship between $X = 1$ and $Y = 1$.

The three definitions discussed so far can be arranged in order of increasing strength. If $\mathbf{X} = \mathbf{x}$ is directly sufficient for $\mathbf{Y} = \mathbf{y}$, then it is also both strongly and weakly sufficient. If $\mathbf{X} = \mathbf{x}$ is strongly sufficient for $\mathbf{Y} = \mathbf{y}$, then it is also weakly sufficient.

## 5.4 A note on Pearl's probability of necessity and sufficiency

For binary treatment $X$ and outcome $Y$, the *probability of necessity* is given by $\mathbb{P}(Y = 0 | do(X = 0), X = 1, Y = 1)$, which is the probability that $Y$ would not have occurred if $X$ did not occur, given that both $X$ and $Y$ actually occurred. Similarly, the *probability of sufficiency* is given by $\mathbb{P}(Y = 1 | do(X = 1), X = 0, Y = 0)$, which measures the capacity of $X$ to produce the outcome $Y$ when starting from absence. This definition is stated to be straightforward to extend to categorical variables in a footnote from Pearl [2009]. Additionally, it is stated in Beckers [2022] that the deterministic counterparts of these definitions turn out to be equivalent to weak sufficiency, though I need to investigate this further.

## 6 Witness Sets

The witness set plays a key role in formalizing the NESS intuition. If $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w})$ is necessary and sufficient for the outcome, and the values $\mathbf{X} = \mathbf{x}$ are themselves necessary and sufficient in the model where $\mathbf{W} = \mathbf{w}$ is fixed, then $\mathbf{X} = \mathbf{x}$ is a necessary element and $(\mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w})$ is the sufficient set.

The inclusion of the witness set reflects a common argument in the philosophy literature, that *parts of causes* themselves deserve to be called causes. The implementation of this idea comes from both the witness set and AC3, the minimality criterion. If we think of $\mathbf{Z} = \mathbf{z}$ as a *complete cause* that is necessary and sufficient for the outcome, if some of its elements $\mathbf{W} \subset \mathbf{Z}$ can be put into the witness set and have their values fixed, and the remaining elements $\mathbf{X} = \mathbf{x}$ can be necessary and sufficient on their own when the values of $\mathbf{W}$, then $\mathbf{X} = \mathbf{x}$ is a part of a cause that can also be called an actual cause in its own right. Since AC3 requires that $\mathbf{X} = \mathbf{x}$ is as small as possible, it does the job of selecting out minimal parts of causes and putting the remaining elements of the complete cause into the witness set.

In our work on actual causality, we have found that in many examples from the literature it is possible to obtain a necessary and sufficient event $\mathbf{X} = \mathbf{x}$ that is a single variable and by definition minimal, and hence there is no requirement for a witness set. However, previous definitions of actual cause before Modified HP did not require the witness to take on actual values $\mathbf{w}^*$, and with these definitions many of the examples in the literature required fixing a witness set in order to satisfy the definition and obtain the correct actual cause.

However, in theory we may still have cases where it is impossible to identify a minimal actual cause without using a witness set. This is illustrated with the following example, due to Caleb Chuck. Say that there is an agent that is shooting a target, and a blocker that will always come in between and defend the target if the agent shoots. Say we are in a scenario where the agent shoots, the blocker activates, and the target is not hit. The activation of the blocker will be an actual cause but the agent shooting will not. This is because the blocker is always activated if the shooter fires and it can never hit the target, but the target would have been hit if the shooter fired and the blocker was not activated. The blocker can hence induce a change in outcome but the shooter cannot. Now say we are in the scenario where the agent does not shoot, the blocker is not activated and the target is not hit. The agent not shooting is not a necessary event because shooting would activate the blocker and the outcome can never change. The blocker is not a necessary event either because in the scenario where the shooter doesn't fire, the outcome does not change whether or not the blocker is activated. However, if we placed the blocker in the witness set, it would be clear that the outcome would change if the shooter fired when the blocker is not activated. Therefore, in this scenario the agent not shooting would be considered an actual cause of the target not being hit, where the absence of the blocker is in the witness set.

Beckers [2021] argues that the sufficiency of events makes no sense if they did not take place, and this is why the witness set is required to take on its actual values $\mathbf{w}^*$. Previous definitions of actual cause before Modified HP did not require the witness to take on actual values $\mathbf{w}^*$, and Beckers

[2021] argues that the sufficiency of events makes no sense if they did not take place. Although it cannot be excluded that the conditions imposed on the variables outside of the witness and $\mathbf{Y}$ somehow ensure the existence of a causally sufficient set, it is not obvious that this is the case or that it will always be possible to find one such set. The way I interpret this is that this definition is an if statement but not only if: although there may be causally sufficient events that don't require only $\mathbf{w}^*$ to be in the witness set, including $\mathbf{w}^*$ in the witness set will give us valid actual causes. That said, there is no discussion of how to choose what can be placed in the witness set; a theory of normality and ranking over counterfactual worlds is included in Halpern [2016], but this is far from complete.

## 7 Normality and Queen of England problems

One criticism that could be leveled at this definition is that it does not distinguish between causes that almost always hold and causes that almost always don't hold. This is illustrated by the following example, due to David Jensen. Say that a rock climber was involved in an accident and survived. The fact that the climber's belt happened to catch on a rock outcropping and the fact that oxygen exists on the surface of planet earth could both have been causes of survival, but they feel very different to human reasoners. An example that has shown up a few times in the philosophy literature is that *the queen of England's not watering my flowers caused my flowers to wilt*, which intuitively seems wrong but is technically correct depending on the definition of actual cause. [Schaffer, 2005] connects this problem to the ambiguity between causation and non-causation, i.e. the non-occurrence of events being causes, but I believe the issue is much more widespread.

Previous work in actual causality has dealt with this using the concept of *normality* [Halpern and Hitchcock, 2015]. Normality is based on intuitions of causal judgements that have been observed in human behavior [Icard et al., 2017, Kominsky et al., 2015, Morris et al., 2018], and is designed to capture the notion of ranking different causes based on strength, plausibility, or intuition. [Halpern and Hitchcock, 2015] proposes a ranking or partial order over all counterfactual worlds, and modifies the definitions of actual cause to incorporate this ranking so that the more *normal* cause is always selected. I believe this is a hacky and unrealistic solution that will not scale to continuous or complex environments.

Our recent work on functional actual causes [Chuck et al., 2023] makes the case for the prevalence of Queen of England problems in ML and RL environments, and motivates this primarily using the example of the agent pushing a block in the presence of an obstacle. What makes this problem particularly challenging is that just limiting the scope of the model (e.g. fewer variables, smaller field of view) and then using the same reasoning does not necessarily make Queen of England problems disappear. This is because Queen of England problems are not just about rare and implausible events, but about *contextually* rare and implausible events. Simply pruning out the entities involved (e.g. the obstacle) from the model is a bad idea since there are other contexts in which they can be really important, and there are way too many contexts to hardcode or store to know ahead of time when an event can be ignored. However, pruning out contextually irrelevant events on the fly is something humans are good at and a good capability for ML algorithms to have. For this reason, we choose to focus on this aspect of actual causation for our ongoing work.

# 8 Examples from the literature

These examples are all taken from different papers and books on actual causality, but most of them are reproduced in Halpern [2016] and Beckers [2021]. Some of these examples also come from the literature on counterfactual harm, and are mostly reproduced from the appendices in Richens et al. [2022]. Like how actual causation aims to capture the NESS intuition from the philosophy literature, the equivalent intuition for harm is the *Counterfactual Comparitive Account (CCA)*, where an event $\mathbf{X} = \mathbf{x}$ harms a person overall if and only if they would have been on balance better off if $\mathbf{X} = \mathbf{x}$ had not occurred. Here, *had not occurred* needs to be formalized using a definition of causation, *if they would have been* using a counterfactual query, *better off* using a utility function, and *on balance* using an aggregation function.

## 8.1 Forest Fires

Suppose that there was a heavy rain in April and electrical storms in the following two months; and in June the lightning took hold. If it hadn't been for the heavy rain in April, the forest would have caught fire in May.

Here $F = 0$ represents the absence of forest fire, $F = 1$ if forest fire occurs in May, and $F = 2$ if forest fire occurs in June. $ES$ represents the presence or absence of electric storms in May and June, and is written as tuple of two binary values. $AS$ expresses whether or not there was rain in April.

$$F = \begin{cases} 2 & \text{if } AS = 1 \wedge ES = (1,1) \\ 1 & \text{if } AS = 0 \wedge (ES = (1,1) \vee ES = (1,0)) \\ 0 & \text{otherwise} \end{cases}$$

The causal graph corresponding to this example is $AS \rightarrow F \leftarrow ES$. The question under consideration is whether $AS = 1$ is a cause of $F = 1 \vee F = 2$, or in other words the rain in April caused a fire. Halpern and Pearl argue that the answer should be no, since the rain only delayed the forest fire from May to June. However we can say that $AS = 1$ is a cause of $F = 2$.

$AS = 1$ alone is not directly sufficient for $F = 2$ since changing $ES$ from $(1,1)$ to anything else will make $F = 0$. Similarly $ES = (1,1)$ is not directly sufficient for $F = 1$ or $F = 2$ since the value of $F$ will flip based on the value of $AS$. This is an example of how direct sufficiency only holds when you cover all the parents.

$(AS = 1, ES = (1,1))$ is directly sufficient for $F = 2$, which means that $AS = 1$ is strongly sufficient for $F = 2$ with $ES = (1,1)$ in the network. However, $(AS = 0, ES = (1,1))$ is directly sufficient for $F = 1$ but not $F = 2$. This shows that there is a contrast value for the intervened variable $AS$, which along with the same setting for the network will not be sufficient for the same outcome $F = 2$. The definition of contrastive necessity given in Beckers [2021] requires the previous statement to hold for all subsets of the network as well, and this works out because $AS = 0$ alone is also not directly sufficient for $F = 2$.

However, we find that it is not possible to have a contrast value of $AS$ that is not strongly sufficient for $F = 1 \vee F = 2$, since we know that $(AS = 0, ES = (1,1))$ is directly sufficient for $F = 1$ and

$(AS = 1, ES = (1,1))$ is directly sufficient for $F = 2$. Hence using strong sufficiency and contrastive necessity agrees with Halpern and Pearl's conclusion.

## 8.2 Rock Throwing

Suzy and Billy both throw a rock at a bottle. Suzy's rock gets there first, shattering the bottle. However Billy's throw was also accurate, and would have shattered the bottle had it not been preempted by Suzy's throw.

Let $BS$ denote whether the bottle shatters or not, $BT$ and $BH$ denote whether Billy threw the rock and whether he hit the bottle respectively, and $ST$ and $SH$ denote whether Suzy threw the rock and whether she hit the bottle respectively.

$$\begin{aligned} SH &= ST \\ BH &= BT \wedge \neg SH \\ BS &= BH \vee SH \end{aligned}$$

This is a classic example of a preemption problem, and is the most commonly studied example in the actual causality literature. The model hardcodes the fact that Suzy's throw gets there first, so all of the definitions correctly judge $ST = 1$ to be a cause of $BS = 1$. For example, $ST = 1$ is strongly sufficient for $BS = 1$ through the network $SH = 1$. However, $ST = 0$ is not strongly sufficient for $BS = 1$ since it is not robust to the value of $BT$; if $BT = 0$ then we would get $BH = 0$ since $\neg SH = \neg ST = 1$ and hence $BS = 0$, but if $BT = 1$ then we would get $BH = 1$ and hence $BS = 1$.

However, weak sufficiency would also lead us to conclude that Billy's throw is a cause of the bottle shattering, or in other words $BT = 1$ would be a cause of $BS = 1$. $BT = 1$ is weakly sufficient for $BS = 1$ under all contexts since the bottle will break for both $ST = 0$ and $ST = 1$ assuming all other variables are set according to the model. Intuitively this means that whether or not Suzy threw, Billy's throw would shatter the bottle if Suzy's did not. However, $BT = 0$ is not weakly sufficient for $BS = 1$ in the context where $ST = 0$, since if both of them did not throw rocks then the bottle is not going to shatter. This contrast can incorrectly lead to the conclusion that Billy's throw is a cause of the bottle shattering.

## 8.3 Is robbery actually harmful?

This is also a preemption problem. Alice robs Bob of his golf clubs. A moment later, Eve would have robbed Bob of his clubs. Therefore, Alice's action does not cause Bob to be worse off as he would have lost his clubs regardless of her actions, and so by the CCA Alice does not harm Bob by robbing him. However, intuitively Alice harms Bob by robbing him, regardless of what occurs later.

### 8.4 Switching railroad tracks

The engineer is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch, so that the train travels down the right-hand track, instead of the left. Since the tracks reconverge up ahead, the train arrives at its destination all the same.

Let $F$ represent the setting of the railroad engineer's switch, and $T$ represent the track along which the train goes. We assume that $F = 0$ makes the train stay on the left side track (i.e. $T = 0$) and $F = 1$ switches it to the right side track (i.e. $T = 0$), hence we have $F = T$. We represent the possibility of a breakdown by setting $T = 2$. The variable $A$ represents whether the train arrives at its destination or not, which will happen if $T \neq 2$.

The causal graph corresponding to this example is $F \to T \to A$. The context is such that $F = 1$, since we know that the engineer flips the switch.

We can see that $F = 1$ is weakly sufficient for $A = 1$, since letting $T$ take on its observational value will naturally lead to $T = 1$ and hence $A = 1$ anyway. $F = 1$ is also strongly sufficient for $A = 1$ with $T$ in the network, since it is directly sufficient for $(T = 1, A = 1)$. However we also have that $F = 0$ is weakly and strongly sufficient for $A = 1$, so according to these definitions both settings of the switch cannot be a cause of the train's arrival. This makes sense since the only thing that can stop an arrival is a breakdown, which according to this model has nothing to do with the switch.

### 8.5 Halt and charge

Major $M$ and sergeant $S$ stand before corporal, and both shout 'Charge!' $(M = 1, S = 1)$. The corporal charges $(C = 1)$. Orders from higher ranking soldiers trump those of lower rank, so if the major had shouted 'Halt' $(M = 0)$ the corporal would not have charged. If the major remains quiet $(M = -1)$, the corporal listens to the sergeant. The majority intuition is that the sergeant did not cause the corporal to charge, because his order was trumped by that of the major.

$$C = \begin{cases} M & \text{if } M \neq -1 \\ S & \text{otherwise} \end{cases}$$

### 8.6 Voting

A ranch has five individuals: $(a_1, ..., a_5)$.They have to vote on two possible outcomes: staying at the campfire $(O = 0)$ or going on a round-up $(O = 1)$. Let $A_i$ denote individual $A_i$'s vote, so $A_i = j$ if individual $a_i$ votes for outcome $j$.

$$O = \begin{cases} A_2 & \text{if } A_1 = A_2 \\ A_1 & \text{if } A_2 = A_3 = A_4 = A_5 \neq A_1 \\ \text{Majority vote} & \text{otherwise} \end{cases}$$

In the actual situation, $A_1 = A_2 = 1$ and $A_3 = A_4 = A_5 = 0$, so the outcome is $O = 1$. We assume that the individuals are not aware of how the voting mechanism works, so that for example they cannot use a vote $A_i = 1$ to flip the outcome from 1 to 0. Halpern and Beckers argue that only $A_1 = 1$ and $A_2 = 1$ should be actual causes of $O = 1$.

### 8.7 Gang leader

An obedient gang is ordered by its leader to join him in murdering someone. All of them, including the leader, shoot the victim at the same time, and the action of any one of them would suffice for the victim's death. Rosenberg and Glymour argue that all of the gang members should individually be considered actual causes, but the Original and Updated HP definitions that were the most current at the time led to the conclusion that only the gang leader's order was the actual cause. Modified HP and all of Beckers' definitions agree with Rosenberg and Glymour.

$$\begin{aligned} Y &= X \vee D \\ X &= D \end{aligned}$$

The context here is $D = 1$. This is a standard case of overdetermination, where $X = 1$ and $D = 1$ are both overdetermining causes of $Y = 1$. The equations can be extended to a logical OR of any number of binary variables that denote whether the corresponding individual shoots, all of which will be equal.

### 8.8 Do we really need counterfactual dependence?

Consider variables $N \in \{0, 1, 2, 3\}$, and binary variables $W, X, Y$. Let $Y = 1$ if and only if $N \neq 0$, and say that:

$$N = \begin{cases} 0 & \text{if } A = 0 \\ 1 & \text{if } A = 1 \wedge X = 1 \\ 2 & \text{if } A = 1 \wedge X = 0 \wedge W = 1 \\ 3 & \text{if } A = 1 \wedge X = 0 \wedge W = 0 \end{cases}$$

The causal graph for this example corresponds to $(A, X, W) \to N \to Y$.

Say we have a context where $A = W = X = 1$. There is no intervention such that $Y = 1$ depends on $X = 1$ under that intervention, since $Y = 1$ will only happen when $A = 1$ anyway, so most of the definitions of causation will not declare $X = 1$ to be a cause of $Y = 1$.

However, $(A = 1, X = 1)$ is strongly sufficient for $Y = 1$ through the network $N = 1$, but $(A = 1, X = 0)$ is not strongly sufficient for $Y = 1$ since the network values $N = 1, Y = 1$ are not robust to the different values of $W$. Hence, by using $A = 1$ as a witness we can declare that $X = 1$ is a cause of $Y = 1$.

### 8.9 Who shot the prisoner?

Suppose that a prisoner dies either if $X$ loads $D$'s gun and $D$ shoots, or if $A$ loads and shoots his gun. Let $Y = 1$ if the prisoner dies and 0 otherwise. Suppose that $X$ loads $D$'s gun ($X = 1$), $D$ does not shoot ($D = 0$), but $A$ does load and shoot his gun ($A = 1$), so that the prisoner dies. Halpern and Pearl argue that clearly $A = 1$ is a cause of $Y = 1$, but also that $X = 1$ should not be a cause of $Y = 1$ since $D$ did not shoot.

### 8.10 Omission problems

*Case 1*: Alice decides not to give Bob a set of golf clubs. Bob would be happy if Alice had given him the golf clubs. Therefore, according to the CCA, Alice's decision not to give Bob the clubs causes Bob harm. However, intuitively Alice has not harmed Bob, but merely failed to benefit him.

*Case 2*: Alice can choose to give Bob her golf clubs or not. She has no obligation to do so. Unbeknownst to her, Eve is planning to rob Bob, but if Bob is holding a golf club she will not dare rob him. Alice decides not to give Bob her golf club and Bob is robbed by Eve. By choosing not to gift her clubs, did Alice harm Bob?

### 8.11 Preventing the worse outcome

*Case 1*: Bob has \$2. The thief Alice is stalking Bob in the marketplace and notices that Eve (a more effective thief) is also stalking Bob. Seeing Eve before Eve notices her, Alice decides to make her move first. She steals \$1 from Bob. Eve was going to steal \$2 from Bob, but is incapable of doing so if someone else robs him first (e.g. Bob realizes he's been robbed and call for the police, making further robbery impossible). Seeing that Bob was robbed by Alice she decides not to rob him.

*Case 2*: In another scenario, Eve has captured Bob and intends to torture him to death. Alice sees this, and is too far away to prevent Eve from doing so. She has a line of sight to Bob (but not Eve) and can shoot him before Eve has a chance to torture him to death, resulting in a painless death.

### 8.12 Do harmful events need to be actual causes?

A soldier falls in a river and is drowning. His officer stands beside the river and has a responsibility to help the soldier if he can. He can choose to swim out to rescue him ($R = T$) or not ($R = F$). There are two enemy sharpshooters watching the river. The first will shoot ($S_1 = T$) if the officer tries to rescue him and wont shoot otherwise, while the second will shoot ($S_2 = T$) if he doesn't try to rescue him, and wont shoot if he does. The officer decides not to rescue the soldier and he is shot dead ($D = T$) by shooter 2. Was the officers decision not to rescue the soldier harmful?

## 9 Implications for future research

The example with the block-pushing agent and the far-away obstacle indicates that a probabilistic and continuous treatment of actual causation should not stop with weak sufficiency and the HP definitions if it is to be useful for studying AI systems and simulation environments. For RL environments it might be possible to get away with direct sufficiency and assume only one-step transition dynamics, but in general it is hard to avoid Queen of England problems with the current definitions. We intend to investigate and extend the strong sufficiency definition to see if it captures what we need, or if we require an alternative definition that lies in between the strong and weak sufficiency definitions.

The definitions presented are extremely difficult to verify in moderately sized environments, as the search for a valid counterfactual assignment that changes the outcome as well as a witness set to fix amounts to a SAT problem. Past work has investigated the use of SAT solvers to discover actual causes. Scalability is a major concern in this line of work, though a continuous definition opens the door to gradient-based approaches that have worked well in other ML problems. Additionally, the datasets used in AI systems are assumed to be both high-dimensional and lying on a lower-dimensional manifold, so a definition that incorporates the idea of *feasible sets* and feasible interventions will be essential for developing scalable approaches.

Some of the key issues in actual causality that need to be addressed in order for it to be widely adopted for applications in ML include:

1. Probabilistic and continuous definitions of actual cause
2. Addressing Queen of England problems and pruning out large sets of unnecessary events
3. A scalable and probabilistic alternative to normality
4. Learning models of the world that can be used to infer actual causes, which is non-trivial with SCMs and CGMs in the presence of context-specific independencies
5. Scalable inference algorithms for discovering actual causes
6. Searching for actual causes in longer chains and general non-Markovian environments

## References

Judea Pearl. *Causality*. Cambridge University Press, New York, 2000.

Joseph Y Halpern. *Actual causality*. MIT Press, 2016.

Sander Beckers. Causal explanations and XAI. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang, editors, *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pages 90–109. PMLR, 11–13 Apr 2022. URL https://proceedings.mlr.press/v177/beckers22a.html.

Hana Chockler and Joseph Y Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

Jonathan Richens, Rory Beard, and Daniel H. Thompson. Counterfactual harm. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 36350–36365. Curran Associates, Inc., 2022.

Sander Beckers, Hana Chockler, and Joseph Halpern. A causal analysis of harm. *Advances in Neural Information Processing Systems*, 35:2365–2376, 2022.

Sander Beckers. Causal sufficiency and actual causation. *Journal of Philosophical Logic*, 50(6): 1341–1374, June 2021. doi: 10.1007/s10992-021-09601-z. URL https://doi.org/10.1007/s10992-021-09601-z.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362, 2021.

Brad Weslake. A partial theory of actual causation. *British Journal for the Philosophy of Science*, 2015.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Jonathan Schaffer. Contrastive causation. *The Philosophical Review*, 114(3):327–358, 2005.

Joseph Y Halpern and Christopher Hitchcock. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 2015.

Thomas F Icard, Jonathan F Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017.

Jonathan F Kominsky, Jonathan Phillips, Tobias Gerstenberg, David Lagnado, and Joshua Knobe. Causal superseding. *Cognition*, 137:196–209, 2015.

Adam Morris, Jonathan Phillips, Thomas Icard, Joshua Knobe, Tobias Gerstenberg, and Fiery Cushman. Judgments of actual causation approximate the effectiveness of interventions. *Psy ArXiv) doi*, 10, 2018.

Caleb Chuck, Sankaran Vaidyanathan, Stephen Giguere, Amy Zhang, David Jensen, and Scott Niekum. Automated discovery of functional actual causes in complex environments. *in submission to the 3rd Conference on Causal Learning and Reasoning*, 2023.