The Counterfactual NESS Definition of Causation

Sander Beckers
Munich Center for Mathematical Philosophy – Ludwig Maximilian University, Munich
srekcebrednas@gmail.com

Abstract

In previous work with Joost Vennekens I proposed a definition of actual causation that is based on certain plausible principles, thereby allowing the debate on causation to shift away from its heavy focus on examples towards a more systematic analysis. This paper contributes to that analysis in two ways. First, I show that our definition is in fact a formalization of Wright's famous NESS definition of causation combined with a counterfactual difference-making condition. This means that our definition integrates two highly influential approaches to causation that are claimed to stand in opposition to each other. Second, I modify our definition to offer a substantial improvement: I weaken the difference-making condition in such a way that it avoids the problematic analysis of cases of preemption. The resulting *Counterfactual NESS definition* of causation forms a natural compromise between counterfactual approaches and the NESS approach.

1 Introduction

Causal models have become very popular tools to offer definitions of *actual causation*, or token causation, in both philosophy and in computer science. Computer scientists (as well as scientists in general) like them because their interventionist semantics fit naturally with the manner in which scientists perform and conceptualize experiments. Philosophers like them because they offer a non-reductive semantics of counterfactuals that avoids the many issues facing a possible-world semantics. It should therefore come as no surprise that the definitions of causation which are expressed using these models fall within the broadly counterfactual approach, which includes interventionist approaches.

Two such definitions have been developed recently by Vennekens and myself (henceforth the *BV definition* (Beckers and Vennekens 2017) and the *BV definition-2* (Beckers and Vennekens 2018)). The proposed definitions by Beckers and Vennekens (BV from now on) distinguish themselves positively from a range of other definitions that are based on that of (Halpern and Pearl 2005) – *HP-style definitions* – by an explicit focus on certain underlying principles that a definition of causation should satisfy, instead of a focus on the – potentially endless – debate regarding many examples and the intuitions they invoke. Both definitions are closely related: the BV definition-2 builds on the simpler BV definition by adding temporal considerations.

The first aim of this paper is to show that the BV definition is in fact a formal integration of the influential *Necessary Element of a Sufficient Set* definition of causation by (Wright 1985, 1988, 2011) into the more popular counterfactual approach. In particular, the BV definition turns out to be equivalent to stating that the NESS definition holds in the *actual* scenario and that the NESS definition does not hold in at least one *counterfactual* scenario. In other words, it combines the NESS approach with a counterfactual difference-making condition. What makes this a surprising result, is that the NESS definition belongs to the regularity approach, which is claimed to stand in opposition to the counterfactual character of causal models. A side-by-side analysis of both approaches reveals that this claim is false.

The second aim of this paper is to modify the BV definitions to overcome two problematic features: they fail to correctly handle standard cases of both *early* and *late* preemption, unless one invokes probabilistic and temporal considerations. By looking at the precise relations between all definitions here considered I propose a weakening of the difference-making condition that solves this problem, resulting in the *Counterfactual NESS* definition of causation, or *CNESS* for short. This analysis shows the CNESS definition to offer a natural compromise between two important approaches to causation.

The paper is structured as follows. The next section introduces the formalism of Structural Equations Modeling that has become widely adopted in the counterfactual approach to causation. Section 3 offers a detailed comparison of the NESS approach and the most popular counterfactual approaches, and then goes on to present a formalization of the NESS definition using structural equations models. Section 4 discusses the three other attempts at formalizing the NESS definition that have been put forward, and shows each of them to be incorrect. Finally, the formalized version of the NESS definition is used in Section 5 to introduce the BV definition of causation and discusses how it relates to HP-style definitions, leading the way to the CNESS definition.

2 Structural Equation Models

This section reviews the definition of causal models as understood in the structural modeling tradition started by

(Pearl 2000). Much of the discussion and notation is taken from (Halpern 2016) with little change.

Definition 1 *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 $\mathcal{R}$ a function that 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*). If $\vec{X} = (X_1, \ldots, X_n)$, $\mathcal{R}(\vec{X})$ denotes the crossproduct $\mathcal{R}(X_1) \times \cdots \times \mathcal{R}(X_n)$. For simplicity, I assume here that $\mathcal{V}$ is finite, as is $\mathcal{R}(Y)$ for every endogenous variable $Y \in \mathcal{V}$.* ∎

Exogenous variables represent factors whose causal origins are outside the scope of the causal model, such as background conditions and noise. The values of the endogenous variables, on the other hand, are causally determined by other variables within the model.

Definition 2 *A* causal model *M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a signature and $\mathcal{F}$ defines a function that associates with each endogenous variable X a* structural equation *F_X giving the value of X in terms of the values of other endogenous and exogenous variables. Formally, the equation F_X maps $\mathcal{R}(\mathcal{U} \cup \mathcal{V} - \{X\})$ to $\mathcal{R}(X)$, so F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$.* ∎

The value of X may depend on the values of only a few other variables. X *depends on* Y if there is some context $\vec{u}$ and a setting of the endogenous variables other than X and Y such that if the exogenous variables have value $\vec{u}$, then varying the value of Y in that context results in a variation in the value of X; that is, there is a setting $\vec{z}$ of the endogenous variables other than X and Y and values y and y' of Y such that $F_X(y, \vec{z}, \vec{u}) \neq F_X(y', \vec{z}, \vec{u})$. We then say that Y is a *parent* of X. $\vec{PA}_X$ denotes all parents of X, which – for reasons of notational simplicity – we also take to include $\mathcal{U}$.

In this paper we restrict attention to *strongly recursive* (or *acyclic*) models, that is, models where, there is a partial order $\preceq$ on variables such that if Y depends on X, then $X \prec Y$. In a strongly recursive model, given a context $\vec{u}$, the values of all the remaining variables are determined (we can just solve for the value of the variables in the order given by $\preceq$). In a strongly recursive model, we often write the equation for an endogenous variable as $X = f(\vec{Y})$; this denotes that the value of X depends only on the values of the variables in $\vec{Y}$, and the connection is given by the function f. For example, we might have $X = Y + 5$.

An *intervention* has the form $\vec{X} \leftarrow \vec{x}$, where $\vec{X}$ is a set of endogenous variables. Intuitively, this means that the values of the variables in $\vec{X}$ are set to the values $\vec{x}$. The structural equations define what happens in the presence of interventions. Setting the value of some 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 $\mathcal{F}$ is replaced by $\mathcal{F}^{\vec{X} \leftarrow \vec{x}}$: for each variable $Y \notin \vec{X}$, $F_Y^{\vec{X} \leftarrow \vec{x}} = F_Y$ (i.e., the equation for Y is unchanged), while for each X' in $\vec{X}$, the equation $F_{X'}$ for X' is replaced by $X' = x'$ (where x' is the value in $\vec{x}$ corresponding to X').

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, an *atomic formula* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$, where

- ϕ is a Boolean combination of atomic formulas,
- $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and
- $y_i \in \mathcal{R}(Y_i)$ for each $1 \leq i \leq k$.

Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\phi$. The special case where $k = 0$ is abbreviated as ϕ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\phi$ 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 *causal setting*, which is a causal model given a context. As usual, we write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in the causal 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 recursive models) solution to the equations in M in context $\vec{u}$ (i.e., the unique vector of values that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). The truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\phi$ if $(M_{\vec{Y} \leftarrow \vec{y}}, \vec{u}) \models \phi$.

3 The NESS Definition of Causation

Given the vital role of causation for assessing liability and guilt, a proper definition of causation is crucial for the legal domain. It has long been acknowledged that the *sine qua non* ('but for') account of causation, which is still officially endorsed in many legal texts, is painfully inadequate as a general definition of causation. In the context of structural equations, this flawed account can be described as equating causation with *counterfactual dependence*.

Definition 3 *Let $(M, \vec{u})$ be a causal setting, let C and E be endogenous variables, and let c and e be values in $\mathcal{R}(C)$ and $\mathcal{R}(E)$ respectively. We say that $E = e$ is* counterfactually dependent *on $C = c$ if $(M, \vec{u}) \models C = c \wedge E = e$ and there exists a $c' \in \mathcal{R}(C)$ such that $(M, \vec{u}) \models [C \leftarrow c']\neg(E = e)$.* ∎

Wright (1985, 1988) has suggested the NESS definition as an alternative definition of causation, in order to deal with the inadequacies of the sine qua non account. (The NESS definition is very similar to Mackie's well-known *INUS* condition. Both are based on the work of (Hart and Honoré 1959).) Over the past few decades this definition has been hotly contested in the literature on causation, culminating in Wright's detailed defense of his definition to the most prominent objections and a discussion on how it differs from the INUS condition (Wright 2011). Central to this defense is the distinction between *lawful sufficiency* and *causal sufficiency*: the former expresses that some effect $E = e$ can be derived from some condition ϕ by a sequence of logical implications, whereas the latter means that $E = e$ can be derived from ϕ by a sequence of instantiations of *causal laws*.

Parallel to the developments in the philosophy of law, Wright's definition also influenced the literature on causation within the formal causal modeling tradition that started with the work of (Pearl 2000). Pearl acknowledges the intuitive appeal of the NESS definition, but criticizes it for lacking precisely those features that Wright appeals to in its defense (Pearl 2009, p. 314): "This basic intuition [the NESS

intuition] is shared by researchers from many disciplines. ... However, all these proposals suffer from a basic flaw: the language of logical necessity and sufficiency is inadequate for explicating these intuitions". Instead, Pearl proposes the language of structural equations, whose semantics intend to capture *causal*, rather than logical, necessity and sufficiency. In other words, Wright and Pearl seem to be in agreement on two fundamental issues, namely that causal sufficiency should be distinguished from logical sufficiency, and that the NESS definition of causation is in the right spirit.

Yet as will become clear below, Pearl's own proposal for defining causation using counterfactuals forms a substantial departure of the NESS definition as understood by Wright, which does not rely on counterfactuals (or at least not to the same extent, more on this later).[1] As a result of this, Wright is "fundamentally opposed to counterfactual analyses of causation and skeptical of attempts by Pearl and others to build formalist accounts based on structural equations."[2] I believe this is a classic case of throwing away the baby with the bathwater: one can perfectly well use structural equations to correctly formalize definitions of causation that do not belong to the counterfactual approach, such as the NESS definition. In order to show this we take a slight detour to look at existing definitions of causation within the structural equations framework.

HP-style Definitions of Causation

The structural equations framework has become a popular formalism for defining actual causation. The definition proposed by Halpern and Pearl (2005) – HP, from now on – has been by far the most influential of the lot. [3] In fact, many of the other definitions using this formalism can be adequately described as offering modifications of the HP definition, and therefore I will refer to them as *HP-style definitions*. [4]

These definitions are often presented with little systematic motivation, focussing instead on capturing the "right" intuition for a myriad of examples. Nevertheless, the reader may verify that the following four assumptions can be used to characterize HP-style definitions.

1. **Formalism** As is clear from the above, they assume that causation can be accurately defined using the structural equations framework. This assumption is rather minimal, but for clarificatory purposes it will turn out to be useful to separate it from three further assumptions that HP-style definitions have in common.

   The second and third assumptions characterize the counterfactual tradition of causation started by (Lewis 1973), which holds that the notions of counterfactual dependence and causation are tightly intertwined:

2. **Dependence** Counterfactual dependence is *sufficient* for causation (but not necessary): if $E = e$ is counterfactually dependent on $C = c$ then $C = c$ causes $E = e$.

3. **Counterfactual** There is also a *necessary* condition for causation that is counterfactual in nature, so that causation always includes a statement about what would have happened had the cause not occurred. This condition can be made precise using Definition 4 introduced later on, for now the following informal version suffices: if $C = c$ causes $E = e$ in some actual scenario then there exists a $c' \neq c$ such that $C = c'$ is not causally sufficient for $E = e$ in the corresponding counterfactual scenario.

4. **Interventionism** They all share the assumption that the relation between counterfactual dependence and causation (roughly) takes on the following form: $C = c$ causes $E = e$ iff $E = e$ is counterfactually dependent on $C = c$ given an intervention $\vec{X} \leftarrow \vec{x}$ that satisfies some conditions P. The divergence between these definitions is to be found in the conditions P that should be satisfied.[5]

I suspect that Wright's scepticism towards using structural equations is based on the observation that these assumptions so often go hand in hand. As will become clear, the BV definition and the CNESS definition show that this need not be the case: both satisfy all of **Formalism**, **Dependence**, and **Counterfactual**, but both reject **Interventionism**, and are thus not HP-style definitions. Therefore opposition to HP-style definitions need not imply opposition to the structural equations framework as such. The family of HP-style definitions form only one particular group of definitions that one can construct using structural equations.

Wright clearly rejects **Counterfactual** and **Interventionism**, but there is nothing in his writings which suggest that he rejects **Dependence**. In fact, as will be argued below, a correct formalization of the NESS definition implies **Dependence** (and falsifies both **Counterfactual** and **Interventionism**). However, in order to get there, it first has to be argued that the NESS definition doesn't conflict with **Formalism**.

The problem here is that the semantics of structural equations are usually given explicitly in terms of counterfactuals, whereas Wright's notion of causal sufficiency is supposedly distinct from a counterfactual interpretation. Rather than settling this issue by getting into the notoriously overloaded concept of a counterfactual, it is more fruitful to simply compare side-by-side the two notions that form the basic building blocks of both approaches.

Structural Equations vs Causal Laws

The structural equations framework as it was developed by Pearl (2009) is not intended to give a reductive account of causation. Its fundamental constituents, the equations, encode basic and autonomous mechanisms that are assumed to be causal themselves. Concretely, these equations encode *scientific laws* (Pearl 2009, p. 27):

[1] Taking my cue from Pearl, elsewhere I develop an alternative formalization of the NESS definition that improves upon Wright's conception (Beckers forthcoming). In many ways the CNESS definition presented here is a simplification of that definition. A detailed comparison is the subject of future work.

[2] Personal communication.

[3] Various versions have been developed by Halpern and Pearl, starting with the one proposed by (Pearl 2000). (Halpern 2016) gives a detailed overview of the different versions.

[4] These are some of the more well-known: (Hitchcock 2001, 2007; Woodward 2003; Hall 2007; Weslake 2015).

[5] (Weslake 2015) offers a detailed analysis of several of these definitions by taking this assumption as his starting point.

> The interpretation of the functional relationship in $Y = f_Y(\vec{X})$ is the standard interpretation that functions carry in physics and the natural sciences; it is a recipe, a strategy, or a *law* specifying what value nature would assign to Y in response to every possible value combination that $\vec{X}$ might take on.[6] [emphasis in original]

This interpretation of a structural equation as a scientific law is entirely analogous to Wright's concept of a *causal law*, which he invokes to explain the distinction between lawful and causal sufficiency (Wright 2011, p. 289):

> A causal law is an empirically derived statement that describes a successional relation between a set of abstract conditions (properties or features of possible events and states of affairs in our real world) that constitute the antecedent and one or more specified conditions of a distinct abstract event or state of affairs that constitute the consequent such that, regardless of the state of any other conditions, the instantiation of all the conditions in the antecedent entails the immediate instantiation of the consequent, which would not be entailed if less than all of the conditions in the antecedent were instantiated.

Moreover, just as a structural equation, Wright's notion of a causal law has a built-in directionality (Ibidem):

> Another critical feature of causal laws – and the related concept of causal sufficiency as distinct from mere lawful sufficiency – is their successional or directional nature, according to which the instantiation of the conditions in the antecedent of the causal law causes the instantiation of the consequent, but not vice versa.

In light of this, there is no a priori reason why the NESS definition could not be formalized using structural equations. We simply need to interpret structural equations as causal laws.

Formalizing the NESS Definition

We now state Wright's NESS definition (Ibidem):

> According to the NESS account as initially elaborated, a condition c was a cause of a consequence e if and only if it was necessary for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of e. The required sense of sufficiency, which ... I call 'causal sufficiency' to distinguish it from mere lawful strong sufficiency, is the instantiation of all the conditions in the antecedent ('if' part) of a causal law, the consequent ('then' part) of which is instantiated by the consequence at issue.

As a first step, we formalize what it means for a set of conditions to be sufficient for the occurrence of a consequent condition in a causal law. For simplicity, we restrict ourselves to consequent conditions that take the form of an atomic formula $E = e$. In a causal model M, the equation F_E contains the combination of all causal laws that together determine the value of E.[7] For example, the equation $E = A \vee B$ states that there are precisely two causal laws which can produce $E = 1$, namely the causal law 'if $A = 1$ then $E = 1$', and the causal law 'if $B = 1$ then $E = 1$'. (As explained earlier, the implication here should not be understood as logical implication.) Therefore "the instantiation of all the conditions in the antecedent ('if' part) of a causal law the consequent ('then' part) of which is instantiated by the consequence at issue", is translated as: the parent variables $\vec{PA}_E$ of E take on values $\vec{pa}_E$ such that $f_E(\vec{pa}_E) = e$.

Definition 4 *Given $\vec{X} \subseteq \mathcal{V}$, we say that $\vec{X} = \vec{x}$ is* sufficient *for $E = e$ w.r.t. $(M, \vec{u})$ if $f_E(\vec{x}, \vec{u}) = e$. (Where $f_E(\vec{x}, \vec{u}) = e$ means that for all settings $\vec{v}' \in \mathcal{V} - \{E\}$ such that the restriction of $\vec{v}'$ to $\vec{X}$ is $\vec{x}$, it holds that $f_E(\vec{v}', \vec{u}) = e$.)* ∎

The following straightforward proposition gives an alternative formulation of sufficiency in terms of interventions.

Proposition 1 *$\vec{X} = \vec{x}$ is sufficient for $E = e$ w.r.t. $(M, \vec{u})$ iff for all values $\vec{y} \in \mathcal{R}(\vec{Y})$ where $\vec{Y} = \mathcal{V} - (\vec{X} \cup \{E\})$, it holds that $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{Y} \leftarrow \vec{y}]E = e$. (Where we assume that $E \notin \vec{X}$.)*

Note that Wright refers to the definition quoted above as the initial version. He offers his preferred version later on, in which the necessity of the cause is made redundant by requiring that the antecedent part of a causal law only states those conditions which are necessary. (Thereby ensuring that any condition c which is part of the antecedent is automatically necessary.) Since structural equations may combine several causal laws into a single equation, I prefer to stick with the first formulation.

Therefore as a second step, we formalize the necessity of the cause "for the sufficiency of a set of existing antecedent conditions that was sufficient for the occurrence of e". That the antecedent conditions exist, simply means that in the actual setting under consideration, those conditions were satisfied. That the cause $C = c$ was necessary for the set to be sufficient, means that the set without the cause is no longer sufficient.

Definition 5 *[Direct NESS] $C = c$* directly NESS-causes *$E = e$ w.r.t. $(M, \vec{u})$ if there exists a $\vec{W} = \vec{w}$ so that the following conditions hold:*

- *$(M, \vec{u}) \models C = c \wedge \vec{W} = \vec{w}$;*
- *$\{C = c, \vec{W} = \vec{w}\}$ is sufficient for $E = e$ w.r.t. $(M, \vec{u})$;*
- *$\vec{W} = \vec{w}$ is not sufficient for $E = e$ w.r.t. $(M, \vec{u})$.*

We call $\vec{W} = \vec{w}$ a witness. ∎

The following straightforward result is useful to clarify what makes this a case of *direct* causation.

Proposition 2 *If $C = c$ directly NESS-causes $E = e$ w.r.t. $(M, \vec{u})$ then there exists a witness $\vec{W} = \vec{w}$ for this such that $\{C = c, \vec{W} = \vec{w}\} \subseteq \vec{PA}_E$.*

[6] I've slightly changed the mathematical notation to make it consistent with the notation used in this paper.

[7] Obviously this means that a structural equation is only an approximate expression of the causal laws, since it only mentions a small subset of the multitude of conditions that a complete specification of the causal laws would require. This is again entirely in line with Wright's conception of causal laws (Ibid, p.290).

Wright (2011) stresses that the NESS definition is distinct from a counterfactual approach because it expresses *weak necessity*, as he calls it, as opposed to *strong necessity*, as he calls counterfactual dependence. $C = c$ is weakly necessary for $E = e$ if there is *a* sufficient condition such that $C = c$ is necessary for its sufficiency, regardless of what other sufficient conditions there might be. A simple application of Definition 5 illustrates this distinction. Say the equation for E is $E = (C \wedge B) \vee (\neg C \wedge A)$, and we are considering a context such that $A = 1$, $B = 1$, and $C = 1$. Then $C = 1$ is a direct NESS-cause of $E = 1$, because it is a necessary element of the sufficient set $\{C = 1, B = 1\}$. The fact that in the counterfactual scenario $C = 0$ becomes sufficient for a different set that would make $E = e$ true – namely $\{C = 0, A = 1\}$ – is entirely irrelevant. Yet it is the existence of that set which explains why $E = 1$ is not counterfactually dependent on $C = 1$.

Although we have now formalized the NESS definition as it is stated explicitly in the above quote, there is a crucial part which is made explicit only in the second version: causation is the transitive closure of the relation we have just defined, i.e., we should also consider a sequence of atomic formulas that satisfy the previous definition.

Definition 6 *[NESS]* $C = c$ NESS-causes $E = e$ *w.r.t.* $(M, \vec{u})$ *if there exists a chain of direct NESS causes from* $C = c$ *to* $E = e$*. (I.e., there exist* $C_1 = c_1$*, . . .,* $C_n = c_n$*, so that* $C = c$ *is a direct NESS cause of* $C_1 = c_1$*, . . ., and* $C_n = c_n$ *is a direct NESS cause of* $E = e$*.)* ∎

I now present a paradigmatic case of early preemption by (Hitchcock 2001, p. 276) to illustrate how the NESS definition works. Preemption cases form the most well-known challenge for any account of causation. Such a case consists of two causal processes that are in competition to produce some effect, and only one of them succeeds.

Example 1 (Backup) *An assassin-in-training is on his first mission. Trainee is an excellent shot: if he shoots his gun, the bullet will fell Victim. Supervisor is also present, in case Trainee has a last minute loss of nerve (a common affliction among student assassins) and fails to pull the trigger. If Trainee does not shoot, Supervisor will shoot Victim herself. In fact, Trainee performs admirably, firing his gun and killing Victim.*

It is intuitively clear that Trainee's shot causes Victim to die. The following equations, with the obvious interpretation of the binary variables, form the standard formalization of this story: $Victim = Trainee \vee Supervisor$ and $Supervisor = \neg Trainee$. The context is such that $Trainee = 1$. (Throughout we follow the standard practice of leaving out the equations for endogenous variables such as $Trainee$ that are determined directly by the context.)

To see that $Trainee = 1$ is a NESS-cause of $Victim = 1$, note that $Trainee = 1$ (or $\{Trainee = 1\}$, to be entirely correct) is sufficient for $Victim = 1$, whereas $\varnothing$ is not. Therefore $Trainee = 1$ directly NESS-causes $Victim = 1$, and a fortiori also NESS-causes $Victim = 1$.

4 Comparison to Other Attempts

As mentioned before, the NESS definition has also been quite influential in the literature on formal approaches to causation. Indeed, this is not the first attempt at formalizing the NESS definition. To the best of my knowledge, three previous attempts have been made, to which we now turn.

Bochman

Bochman (2018) offers the most recent attempt at formalizing the NESS definition. In fact, he was the first to suggest there is a connection to the work of (Beckers and Vennekens 2018).[8] However, there are two major problems with his own attempt that conclusively show it to be incorrect: contrary to the NESS definition, his definition of causation is not transitive, and it does not satisfy **Dependence**.[9]

Halpern

Halpern (2008) is the latest to offer an attempt at formalizing the NESS definition in the structural equations framework. Unfortunately, just as the attempt discussed below, it was developed before Wright's substantial clarifications in his 2011 article, and therefore it shouldn't come as a surprise that both of them miss the mark. There are two problems with his attempt, only one of which is conclusive evidence that his definition is not a correct formalization of the NESS test. Rather than presenting his definition, I simply state the problems, and leave it to the reader to verify that these apply.

The first problem is only a potential problem, which requires further investigation to be settled: it is not clear whether his definition is transitive. If it is not, then just as with Bochman's definition, it conclusively fails as a formalization of the NESS test. Although I was unable to find any violations of transitivity, Halpern does not mention transitivity and nothing in his formulation suggests that he intended his definition to be transitive.[10] The first problem does not allow us to reach a conclusive verdict, but the second one does: Halpern's definition violates **Dependence** (and the NESS definition does not, see Section 5.) Consider again Example 1, but looking at the context in which $Trainee = 0$, and thus it is Supervisor who shoots Victim instead. In that case, $Supervisor = 1$ directly NESS-causes $Victim = 1$ (also, $Victim = 1$ counterfactually depends on $Supervisor = 1$). Yet, as the reader may verify, Halpern's definition does not consider $Supervisor = 1$ a cause of $Victim = 1$.

Baldwin and Neufeld

Baldwin and Neufeld (2004) were the first to suggest a formalization of the NESS definition in the structural equations framework. Their attempt faces two problems.

[8] It was reading that remark in his paper that provided me with the impetus for writing this one.

[9] Concretely, Bochman writes: "In our theory, general causal inference is transitive, while actual causation is not." Also, his Example 8 is a case of counterfactual dependency without causation. (Bochman 2018, p. 1735)

[10] In personal communication he has confirmed that he did not intend his definition to be transitive and has not looked into whether it is.

The first problem runs deeper than just the discussion about the NESS definition: their definition makes use of syntactic properties of a structural equation, which is at odds with the manner in which structural equations are usually interpreted (Pearl 2009, p. 314-315). To illustrate this point, consider the equation $E = A \wedge B$ and the context in which both $A = 0$ and $B = 0$. In this case, the NESS definition judges both $A = 0$ and $B = 0$ to be causes of $E = 0$. However, Baldwin and Neufeld's definition judges neither to be causes. But if we write the equation as $\neg E = \neg A \vee \neg B$, then their definition does judge both to be causes. However, both equations are semantically equivalent, and there is no indication in Wright's work that he considers such syntactic properties to be relevant.

The second problem is that their definition is an HP-style definition, because it satisfies all four assumptions discussed in Section 3. As was noted before, this is inconsistent with Wright's views of the NESS definition.

5 The BV Definition of Causation

All of Definitions 4, 5, and 6, appear explicitly in the work of BV, be it under different names (Beckers and Vennekens 2017, 2018). A direct NESS cause is called "a direct actual contributor" and a NESS cause is called "an actual contributor". BV do not mention a connection to the NESS definition, for the simple reason that we were unaware such a connection existed.

The BV definition can now be formulated as follows.

Definition 7 *[BV definition]* $C = c$ BV-causes $E = e$ w.r.t. $(M, \vec{u})$ *if* $C = c$ *NESS-causes* $E = e$ *w.r.t.* $(M, \vec{u})$ *and there exists a* $c' \in \mathcal{R}(C)$ *such that* $C = c'$ *does not NESS-cause* $E = e$ *w.r.t.* $(M_{C \leftarrow c'}, \vec{u})$. ∎

Several observations are apparent immediately. The first is that being a NESS-cause is a necessary condition for being a BV-cause. BV even go as far as claiming that when we restrict attention to binary variables, this also holds for all of the definitions that we have been referring to as HP-style definitions. Although the claim is true for many examples, the following example shows that it is false in general.

We have as equations $E = (C \wedge D) \vee A$ and $A = \neg D$. Now consider a context such that $D = 0$ and $C = 1$, and thus also $A = 1$ and $E = 1$. It is easy to see that $C = 1$ is not a NESS-cause of $E = 1$: $A = 1$ is sufficient by itself, and $\{C = 1, D = 0\}$ is not sufficient for $E = 1$. Yet most of the HP-style definitions do consider $C = 1$ to be a cause of $E = 1$. Recall that **Interventionism** demands $E = 1$ counterfactually depends on $C = 1$ under some intervention, and this intervention satisfies certain conditions P. The intervention $[D \leftarrow 1]$ clearly meets the first requirement. As details of the conditions P differ between the definitions, we focus on the standard HP definition itself. It demands that $E = 1$ holds in this context under all interventions that consist of $C = 1$, any subset of the chosen intervention $[D \leftarrow 1]$, and any subset of the intervention that sets the other variables to their actual value, i.e., $E = 1$ has to hold in this context for $[C \leftarrow 1]$, $[C \leftarrow 1, D \leftarrow 1]$, $[C \leftarrow 1, A \leftarrow 1]$, $[C \leftarrow 1, D \leftarrow 1, A \leftarrow 1]$. This demand is met.

The following result from (Beckers and Vennekens 2017) allows us to make two more observations.

Theorem 1 *$E = e$ is counterfactually dependent on $C = c$ w.r.t. $(M, \vec{u})$ iff $C = c$ NESS-causes $E = e$ w.r.t. $(M, \vec{u})$ and there exist $c' \in \mathcal{R}(C)$, $e' \neq e \in \mathcal{R}(E)$, such that $C = c'$ NESS-causes $E = e'$ w.r.t. $(M_{C \leftarrow c'}, \vec{u})$.*

This result implies a second observation: both the BV definition and the NESS definition satisfy **Dependence**. Yet neither definitions are HP-style definitions. What makes the NESS definition distinct from HP-style definitions is that it does not satisfy **Counterfactual** or **Interventionism**.

Theorem 2 *NESS satisfies* **Dependence***, but does not satisfy* **Counterfactual** *or* **Interventionism**.

Proof: Dependence is a direct consequence of Theorem 1.

Counterfactual and **Interventionism**. Say we have the following equations: $E = D \vee \neg C$ and $D = C$. Consider the context such that $C = 1$. Then, according to any definition that accepts **Dependence**, it holds that $C = 1$ causes $D = 1$ and $D = 1$ causes $E = 1$. By transitivity, $C = 1$ NESS-causes $E = 1$. Nevertheless, we also have that $C = 0$ is sufficient for $E = 1$ in the counterfactual setting $[C \leftarrow 0]$, thus violating **Counterfactual**. Further, since D is the only variable besides C and E, the only two interventions that are possible candidates for satisfying **Interventionism** are $D \leftarrow 1$ and $D \leftarrow 0$, neither of which work. ∎

The BV definition on the other hand is distinct from HP-style definitions only because it does not satisfy **Interventionism**.

Theorem 3 *BV satisfies* **Dependence** *and* **Counterfactual***, but does not satisfy* **Interventionism**.

Proof: Dependence Assume that $E = e$ is counterfactually dependent on $C = c$ w.r.t. $(M, \vec{u})$. Then, by Theorem 1, we know that $C = c$ NESS-causes $E = e$ w.r.t. $(M, \vec{u})$ and there exists a $c' \in \mathcal{R}(C)$, $e' \neq e \in \mathcal{R}(E)$, such that $C = c'$ NESS-causes $E = e'$ w.r.t. $(M_{C \leftarrow c'}, \vec{u})$. From the latter it follows that $(M_{C \leftarrow c'}, \vec{u}) \models E = e'$, and thus $C = c'$ does not NESS-cause $E = e$ w.r.t. $(M_{C \leftarrow c'}, \vec{u})$.

Counterfactual Assume that $C = c$ BV-causes $E = e$ w.r.t. $(M, \vec{u})$. Say $c' \in \mathcal{R}(C)$ is such that $C = c'$ does not NESS-cause $E = e$ w.r.t. $(M_{C \leftarrow c'}, \vec{u})$. We proceed by a reductio: assume that $C = c'$ is sufficient for $E = e$ w.r.t. $(M_{C \leftarrow c'}, \vec{u})$. By Definition 5, this means that either $C = c'$ is a direct NESS-cause of $E = e$ w.r.t. $(M_{C \leftarrow c'}, \vec{u})$, or that $\varnothing$ is sufficient for $E = e$ w.r.t. $(M_{C \leftarrow c'}, \vec{u})$. The former contradicts our assumption, so it has to be the latter. In other words, $f_E(\vec{u}) = e$. But this implies that $E = e$ has no direct NESS causes, and thus also no NESS-causes. This contradicts the assumption that $C = c$ BV-causes $E = e$.

Interventionism Assume we have a model with the following equations (all variables are binary): $E = F \vee \neg A \wedge \neg D$, $F = D$, and $D = C \vee A$. Say the context is such that $A = 1$ and $C = 1$ hold. Then clearly $C = 1$ BV-causes $E = 1$. Yet there is no intervention such that $E = 1$ counterfactually depends on $C = 1$ given that intervention. To see why, note that such an intervention could not include D or F, for any influence of C on E goes through those variables. So the only candidates are $A \leftarrow 1$ and $A \leftarrow 0$, neither of which work. ∎

A third observation is that both the BV definition and counterfactual dependence consist of the NESS definition combined with a counterfactual difference-making condition. Concretely, they both require that there exists a counterfactual value which fulfills a different role than the actual value. In the case of counterfactual dependence, that different role is to NESS-cause a counterfactual value of the effect. The BV definition merely demands that the counterfactual value does not fulfill the same role as the actual value, without further specification. Thus the BV definition represents a compromise between counterfactual dependence and NESS-causation: it adds a counterfactual component to NESS-causation that guarantees causes make a difference to their effect, but does not demand that this difference consists in preventing the effect from occurring.

6 The CNESS Definition of Causation

Consider again Example 1. We already concluded that $Trainee = 1$ NESS-causes $Victim = 1$. To assess whether $Trainee = 1$ also BV-causes $Victim = 1$, we have to look at the counterfactual scenario that is the result of the intervention $[Trainee \leftarrow 0]$. In that scenario, $Trainee = 0$ directly NESS-causes $Supervisor = 1$, which directly NESS-causes $Victim = 1$, leading to the conclusion that $Trainee = 0$ also NESS-causes $Victim = 1$. So Victim dies either way. More important for our purposes (but probably not for Victim) is that the candidate cause does not meet BV's difference-making condition, and thus $Trainee = 1$ does not BV-cause $Victim = 1$.

BV show in detail that the same verdict is reached for other examples of early preemption (Beckers and Vennekens 2017). Although BV attempt to justify this verdict by appealing to probabilistic considerations, there is a cheaper – and arguably more compelling – solution: adopt a more subtle counterfactual difference-making condition. The idea is to capture the role of some event $C = c$ in a more nuanced way than merely stating that it NESS-causes an effect $E = e$, by looking at the specific path along which the chain of direct NESS causes proceeds.

Definition 8 $C = c$ NESS-causes $E = e$ along a path p *w.r.t.* $(M, \vec{u})$ *if the values of the variables in* p *form a chain of direct NESS causes from* $C = c$ *to* $E = e$. *(I.e., there exists a path* $p = (C_1, \ldots, C_n)$*, and values* $c_1, \ldots, c_n$*, so that* $C = c$ *is a direct NESS cause of* $C_1 = c_1$*, . . ., and* $C_n = c_n$ *is a direct NESS cause of* $E = e$.*)* ∎

By focussing on a specific path and its subpaths, it becomes possible to formulate a more subtle difference-making condition. Say $C = c$ NESS-causes $E = e$ along a path p, and there is some $C = c'$ that would also NESS-cause $E = e$ along some path q (and not along any other path). If q contains variables that do not appear in p, then $C = c$ and $C = c'$ do not fulfill the same causal role, and thus $C = c$ made a difference to the manner that the effect $E = e$ came about. Implementing this idea results in the *Counterfactual NESS definition*, which I suggest as an improvement of the BV definition.

Definition 9 *[CNESS definition]* $C = c$ CNESS-causes $E = e$ w.r.t. $(M, \vec{u})$ *if* $C = c$ *NESS-causes* $E = e$ *along some path* p *w.r.t.* $(M, \vec{u})$ *and there exists a* $c' \in \mathcal{R}(C)$ *such that* $C = c'$ *does not NESS-cause* $E = e$ *along any subpath* p' *of* p *w.r.t.* $(M_{C \leftarrow c'}, \vec{u})$. *(A subpath of* p *is a path whose variables are all members of* p.*)* ∎

Theorem 4 *CNESS satisfies* **Dependence** *and* **Counterfactual***, but does not satisfy* **Interventionism**.

Proof: The proof of Theorem 3 applies without change. (Note that direct NESS-causes are NESS-causes along the empty path.) ∎

Applying this definition to Example 1, we see that $Trainee = 1$ NESS-causes $Victim = 1$ directly (i.e., along the empty path), whereas $Trainee = 0$ NESS-causes $Victim = 0$ along $Supervisor$. As a result, $Trainee = 1$ CNESS-causes $Victim = 1$.

Lastly, although I have refrained from discussing the temporal considerations that BV invoke to deal with cases of Late Preemption, I point out that the CNESS definition handles standard cases of Late Preemption as the following one without invoking such temporal considerations, suggesting that it also offers an alternative to the BV definition-2 (Beckers and Vennekens 2018).

Example 2 *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.*

(Halpern and Pearl 2005) use the following variables for this example, which capture the fact that Billy's throw was preempted by Suzy's rock hitting the bottle: BS for the bottle shattering, BH, SH for Billy's (resp. Suzy's) rock hitting the bottle, and two more variables (BT, ST) for either of them throwing their rock. The equations are then as follows: $BS = BH \vee SH$, $SH = ST$, $BH = BT \wedge \neg SH$. The reader may verify that in the context under consideration in which $ST = 1$ and $BT = 1$, the BV definition offers the mistaken verdict that $ST = 1$ does not cause $BS = 1$, whereas the CNESS definition does not. (Note also that in the scenario where Suzy *does not throw*, the NESS definition judges $ST = 0$ to be a cause of $BS = 1$. This is an unacceptable result that the BV and CNESS definitions avoid.)

7 Conclusion

I have formalized Wright's NESS definition of causation using the framework of structural equations modeling, thereby showing that this framework is not exclusively suitable for counterfactual approaches. Moreover, this formalization revealed that the recent approach by Beckers & Vennekens offers a nice compromise between a regularity approach and the most popular counterfactual approaches based on that of Halpern & Pearl. Further spelling out the relations between all these definitions allowed a modification of the BV definition that removes a major disadvantage of that approach. The resulting Counterfactual NESS definition of causation combines the NESS definition with a subtle counterfactual difference-making condition. In future work I intend to explore further properties of this definition.

Acknowledgments

Many thanks to Hein Duijf, Joe Halpern, Joost Vennekens, Marc Denecker, and AAAI reviewers for comments on earlier versions of this paper. This research was made possible by funding from the Alexander von Humboldt Foundation.

References

Baldwin, R. A.; and Neufeld, E. 2004. The Structural Model Interpretation of the NESS Test. In *Advances in Artificial Intelligence*, volume 3060, 297–307. Lecture Notes in Computer Science.

Beckers, S. forthcoming. Causal Sufficiency and Actual Causation. *Journal of Philosophical Logic* URL https://arxiv.org/abs/2102.02311.

Beckers, S.; and Vennekens, J. 2017. The Transitivity and Asymmetry of Actual Causation. *Ergo* 4(1): 1–27.

Beckers, S.; and Vennekens, J. 2018. A Principled Approach to Defining Actual Causation. *Synthese* 195(2): 835–862.

Bochman, A. 2018. Actual Causality in a Logical Setting. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 1730–1736.

Hall, N. 2007. Structural equations and causation. *Philosophical Studies* 132(1): 109–136.

Halpern, J.; and Pearl, J. 2005. Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science* 56(4): 843–87.

Halpern, J. Y. 2008. Defaults and Normality in Causal Structures. In *Principles of Knowledge Representation and Reasoning: Proc. Eleventh International Conference (KR'08)*, 198–208.

Halpern, J. Y. 2016. *Actual Causality*. MIT Press.

Hart, H.; and Honoré, T. 1959. *Causation in the Law*. Oxford University Press.

Hitchcock, C. 2001. The intransitivity of causation revealed in equations and graphs. *Journal of Philosophy* 98: 273–299.

Hitchcock, C. 2007. Prevention, Preemption, and the principle of Sufficient Reason. *The Philosophical review* 116(4): 495–532.

Lewis, D. 1973. Causation. *Journal of Philosophy* 70: 113–126.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. ISBN 0-521-77362-8.

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

Weslake, B. 2015. A Partial Theory of Actual Causation. *The British Journal for the Philosophy of Science* forthcoming.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press. ISBN 9780195155273.

Wright, R. W. 1985. Causation in Tort Law. *California Law Review* 73: 1735–1828.

Wright, R. W. 1988. Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts. *Iowa Law Review* 73: 1001–1077.

Wright, R. W. 2011. The NESS Account of Natural Causation: A Response to Criticisms. In Goldberg, R., ed., *Perspectives on Causation*. Hart Publishing.