The Structural Model Interpretation of the NESS Test

Richard A. Baldwin and Eric Neufeld

Department of Computer Science, University of Saskatchewan,
Saskatoon, Canada SK S7H 3A8
{rab983,eric}@cs.usask.ca

Abstract

Within the law, the traditional test for attributing causal responsibility is the counterfactual "but-for" test, which asks whether the injury complained of would have occurred but for the defendant's wrongful act. This definition generally conforms to common intuitions regarding causation, but gives non-intuitive results in situations of overdetermination with two or more potential causes present. To handle such situations, Wright defined the NESS Test of causal contribution, described as a formalization of the concept underlying common intuitions of causal attribution. Halpern and Pearl provide a definition of actual causality in the mathematical language of structural models that yields counterintuitive results in certain scenarios. We present a new definition that appears to correct those problems and explain its greater conformity with the intuitions underlying the NESS test.

Introduction

According to Ashley (1990, p. 2), the legal domain is of interest to AI because it is between those formal domains so amenable to knowledge representation and those commonsense domains whose representation remains so elusive. Here we discuss *actual causation* (also called *causation in fact*) in the law from both perspectives. In common law jurisdictions, actual causation must be established between the defendant's wrongful conduct and the injury suffered for liability to be attached. The generally accepted test for determination of actual causation in the law is a counterfactual test, the *but-for* test. If the specified injury would not have occurred 'but for' the defendant's wrongful conduct, then actual causation is established. The test is considered straightforward enough to be applied by juries. However, the test is not comprehensive. It is known to fail when the scenario describing the injury includes other potential causes that would have caused the specified injury in the absence of the defendant's wrongful conduct. This is known as *overdetermined* causation.

Wright (1975, pp. 1775-76) divides cases of overdetermined causation into preemptive and duplicative causation cases. In preemptive causation, the effect of other potential causes is preempted by the effect of the defendant's wrongful act. For example, the defendant stabs and kills the victim before a fatal dose of poison previously administered by a third party can take effect. In duplicative causation, the effect of the defendant's act combines with, or duplicates, the effect of other potential causes where the latter were alone sufficient to bring about the injury. For example, the defendant and another party start separate fires that combine to burn down the victim's house where each fire was independently sufficient to do so. Since in these cases it is not true that 'but for' the defendant's wrongful act the specified harm would not have occurred, according to the but-for test, in neither scenario is the defendant's conduct an actual cause of the injury. Such a result is contrary to intuitions about responsibility and, by implication, about causality.

To cope with overdetermination, Wright (1985) proposes a comprehensive test for actual causation, the NESS (Necessary Element of a Sufficient Set) test. He adopts the view that there is an intelligible, determinate concept of actual causation underlying and explaining common intuitions and judgments about causality and that this concept explains the "intuitively plausible factual causal determinations" of judges and juries when "not confined by incorrect tests or formulas." Wright (1985, p. 1902) contends that not only does the NESS test capture the common-sense concept underlying these common intuitions and judgments, the NESS test defines that concept.

Pearl (2000, pp. 313-15) claims that while the intuitions underlying the NESS test are correct the test itself is inadequate to capture these intuitions because it relies on the traditional logical language of necessity and sufficiency, which cannot capture causal concepts. Pearl (Pearl, 1995; Galles and Pearl, 1997; Galles and Pearl, 1998; Pearl, 2000) proposes a mathematical language of structural causal models (structural language) for formalizing counterfactual and causal concepts. Pearl (1998; 2000, Chap. 10) first applies this structural language to define actual causation using a complex construction called a causal beam. (Halpern and Pearl, 2000) develops a "more transparent" definition (Halpern-Pearl definition), but still using structural models.

In (Baldwin and Neufeld, 2003) we suggested that the Halpern-Pearl definition essentially formalizes Wright's NESS test. However, a result of Hopkins and Pearl (2003) shows that this is not the case. In the sequel we discuss the implications of the Hopkins and Pearl result for the relationship between the Halpern-Pearl definition and the

NESS test and, in response, we present an alternative structural language definition of actual causation which we believe does capture the essential meaning of the NESS test. We illustrate this through re-analysis of the examples in (Baldwin and Neufeld, 2003) and in the process lend validity to Wright's controversial NESS analysis of a complex class of causal scenarios known as double omission cases.

The Ness Test

Wright (1985, 1988, 2001) describes the NESS test as a refinement and development of the concept of a causally relevant condition as developed by Hart and Honore (1985). As Wright describes their analysis, this concept of singular (actual) causation depends on the core concept of general causality we all employ that conforms to and is explained by a regularity account of causation attributed to Hume as modified by Mill. To Hume is attributed the idea that causal attributions exhibit a belief that a succession of events fully instantiates some causal laws or generalizations (incompletely described causal laws). A causal law is described as an if-then statement whose antecedent lists minimal sufficient sets of conditions for (necessary for the sufficiency of) the consequent (the effect). Mill's contribution to the analysis is that there may be more than one set of sufficient conditions for an effect in general and in particular situations (the "plurality of causes" doctrine). Stated in full, the NESS test requires that a particular condition was a cause of (condition contributing to) a specific consequence if and only if it was a necessary element of a set of antecedent actual conditions that was sufficient for the occurrence of the consequence.

In circumstances where only one actual or potential set of conditions is sufficient for the result, the NESS test reduces to the but-for test (Wright, 1985). To illustrate that the NESS test matches common intuitions where the but-for test fails, Wright considers three variations of a two-fire scenario: fires X and Y are independently sufficient to destroy house H if they reach it and they are the only potential causes of house H's destruction so that if neither reach the house it will not be destroyed. In the first situation, X reaches and destroys H and Y would not have reached H even if X were absent. The common intuition here is that X was a cause of the destruction of H but not Y. In this case there is a single actually sufficient set of conditions and no other even potentially sufficient set of conditions. (This assumes that actually sufficient sets of conditions are minimal.) X was a necessary element (necessary for the sufficiency) of that single, actually sufficient set, a NESS condition. It was also a but-for condition.

In the second situation, X and Y reach H simultaneously and combine to destroy it. Here Wright claims that the common intuition is that both X and Y were causes of the destruction of the house. There are two overlapping sets of actually sufficient conditions. X is necessary for the sufficiency of the set including itself but not Y and Y is necessary for the sufficiency of the set including itself but not X. Neither X nor Y is a but-for cause of the destruction of H but each is a duplicative NESS cause.

In the final situation, X reaches and destroys H before Y can arrive and, if X had been absent, Y would have destroyed H. Here the common intuition is unquestionably that X caused the destruction of H and Y did not. Fire Y is not a NESS condition for the destruction of H since any actually sufficient set of conditions, given the assumptions of the scenario, must include X, and Y is not necessary for the sufficiency of any set of conditions that includes X. Fire X, on the other hand, is necessary for the sufficiency of the actually sufficient set of which it is a member. Because the set containing Y but not X would have been sufficient in the absence of X, X is not a but-for cause of the destruction of H. X was a preemptive NESS cause because it preempted the actual sufficiency of the potentially sufficient set including Y.

The Structural Equation Model

Following (Halpern and Pearl, 2001; Pearl, 2000) a *signature* $\mathcal{S}$ is a 3-tuple ($\mathcal{U}$, $\mathcal{V}$, $\mathcal{R}$), where $\mathcal{U}$ is a finite set of exogenous variables, $\mathcal{V}$ is a set of endogenous variables, and $\mathcal{R}$ is a relation associating with each variable $\mathrm{Y} \in \mathcal{U} \bigcup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (the *range* of Y).

A *causal model* over a signature $\mathcal{S}$ is a 2-tuple $M = (S, F)$, where $\mathcal{F}$ is a relation associating each $X \in \mathcal{V}$ with a function F_X that describes the outcome of X given the values of other variables in the model. These functions are the structural equations and describe the process by which the dependent variable receives its value; they correspond to causal mechanisms (law-like regularities) in the domain being modeled. The values of endogenous variables are determined by the values of other endogenous variables and exogenous variables. The values of exogenous variables are determined outside the model and are given. A particular setting $\vec{u}$ of variables in $\mathcal{U}$ is a *context* for the model M and a model with a given context is a *causal world*.

We consider *recursive* causal models, where for any two variables X and Y either F_X is independent of the value of Y (i.e., $F_X(\ldots, y, \ldots) = F_X(\ldots, y', \ldots)$ for all $y, y' \in \mathcal{R}(Y)$) or F_Y is independent of the value of X. Recursive causal models have a unique solution, a unique set of variable values simultaneously satisfying all model equations. If PA_X is the minimal set of variables in $\mathcal{V} - X$ and U_X the minimal set of variables in $\mathcal{U}$ that together suffice to represent F_X, a recursive causal model gives rise to a *causal diagram*, a directed acyclic graph (DAG) where each node corresponds to a variable in $\mathcal{V} \bigcup \mathcal{U}$ and the directed edges point from members of $PA_X \bigcup U_X$ to X and are *direct causes* of X. (PA_X connotes the *parents* of X, conventionally restricted to endogenous variables.) The edges in a causal diagram represent the non-parameterized (or arbitrary) form of the function for a variable, $X = F_X(U_X, PA_X)$.

An external *intervention* (or *surgery*) setting $X = x$, where $X \in \mathcal{V}$, is denoted $X \leftarrow x$ and amounts to pruning the equation for X from the model and substituting $X = x$ in the remaining equations. An intervention that forces the values of a subset of $\mathcal{V}$ prunes a subset of equations from the model, one for each variable in the set, and substitutes the corresponding forced values in the remaining equations. (A set X of variables in $\mathcal{V}$ is sometimes written $\vec{X}$ and a setting of those vectors is written $\vec{X} \leftarrow \vec{x}$.) Interventions represent non-modeled contingencies perturbing causal mechanisms. The result of an intervention $\vec{X} \leftarrow \vec{x}$ is a new causal model (a *submodel*), denoted $M_{\vec{X} \leftarrow \vec{x}}$, over the signature $\mathcal{S}_{\vec{X}} = (\mathcal{U}, \mathcal{V} - \vec{X}, \mathcal{R}|_{\mathcal{V}-\vec{X}})$. In the corresponding causal diagram, it amounts to removing the edges from $PA_X \cup U_X$ to X. A submodel represents a counterfactual world.

For a given signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, where $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *basic causal formula* is of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where φ is a Boolean combination of primitive events, $Y_1 \ldots Y_k$ are distinct variables in $\mathcal{V}$, and $y_i \in \mathcal{R}(Y_i)$. Basic causal formulas are abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$ or just φ when $k = 0$. A *causal formula* is a Boolean combination of basic causal formulas.

A basic causal formula is true or false in a causal model given a context $\vec{u}$. Where ψ is a causal formula (or a Boolean combination of primitive events), $(M, \vec{u}) \models \psi$ means that ψ is true in the causal model M in the context $\vec{u}$. $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](X = x)$ means that X has value x in the unique solution to the equations in the submodel $M_{\vec{Y} \leftarrow \vec{y}}$ in context $\vec{u}$. In other words, in the world in which $\mathcal{U} = \vec{u}$, the model predicts that if $\vec{Y}$ had been $\vec{y}$ then X would have been x; that is, in the counterfactual world $M_{\vec{Y} \leftarrow \vec{y}}$, resulting from the intervention $\vec{Y} \leftarrow \vec{y}$, X has the value x. Causes are conjunctions of primitive events of the form written $\vec{X} = \vec{x}$.

Definition (*Actual Cause; Halpern-Pearl*): $\vec{X} = \vec{x}$ is an *actual cause* of φ in a model M in the context $\vec{u}$ (i.e., in $(M, \vec{u})$) if the following conditions hold:

HC1. $(M, \vec{u}) \models (\vec{X} = \vec{x}) \wedge \varphi$.

HC2. There exists a partition $(\vec{Z}, \vec{W})$ of $\mathcal{V}$ with $\vec{X} \subseteq \vec{Z}$ and some setting $(\vec{x}', \vec{w}')$ of the variables in $(\vec{X}, \vec{W})$ such that, where $(M, \vec{u}) \models Z = z^*$ for each $Z \in \vec{Z}$ (i.e., the actual value of z in context $\vec{u}$

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

(b) $(M, \vec{u}) \models [\vec{X} \leftarrow \vec{x}, \vec{W} \leftarrow \vec{w}', \vec{Z}' \leftarrow \vec{z}^*]\varphi$ for every subset $\vec{Z}'$ of $\vec{Z}$.

HC3. $\vec{X}$ is minimal; no subset of $\vec{X}$ satisfies conditions HC1 and HC2.

Condition HC2 exemplifies what has been called the "counterfactual strategy" for defining actual causation—"event C causes event E *iff* for some *appropriate* G, E is counterfactually dependent on C when we hold G fixed" (Hopkins and Pearl, 2003). The basic idea is that an active causal process (a set Z satisfying condition HC2) be shielded from spurious (preempted) or redundant (duplicative) causal processes before testing whether the effect is counterfactually dependent on the putative cause. However, contrary to the claim in (Baldwin and Neufeld, 2003) that the Halpern-Pearl definition essentially formalizes Wright's NESS test, it turns out that the choice of "appropriate G" under condition HC2 is too permissive to represent the NESS test in the language of structural models.

The Halpern-Pearl Definition and NESS

Hopkins and Pearl (2003) show that even locally, between a variable (the effect) and its parents (direct causes), the Halpern-Pearl definition does not require that an active causal process (a set Z) be actually sufficient; counterfactual dependence satisfying condition HC2(b) may depend on *non-actual* conditions.

To see this, consider a causal model M with context $\mathcal{U} = \vec{u}$. For simplicity, assume that all $U \in \mathcal{U}$ are trivial in the structural equation F_X for X (i.e., U_X is empty) so that $F_X : Dom(PA_X) \rightarrow Dom(X)$. Now consider an assignment $\overrightarrow{PA_X} \leftarrow \overrightarrow{pa_X}$ (here, the vector notation represents an ordered assignment of values $\overrightarrow{pa_X}$ to variables $\overrightarrow{PA_X}$) such that $X = x$. If $\overrightarrow{PA_X} = \{V_1, \ldots, V_n\}$ and $\overrightarrow{pa_X} = \{v_1, \ldots, v_n\}$, the logical sentence consisting of the conjunction of literals $V_i = v_i$ (i.e., $V_1 = v_1 \wedge \ldots \wedge V_n = v_n$) implies $X = x$. If such a sentence is formed for each assignment $\overrightarrow{PA_X} \leftarrow \overrightarrow{pa_X}$ such that $X = x$ and a new sentence, denoted $\Delta(X = x)$, is formed as a disjunction of all such sentences (so that the resulting logical sentence is in disjunctive normal form), then $X = x$ *iff* $\Delta(X = x)$. To illustrate, Hopkins and Pearl give this example.

Example (firing squad) A firing squad consists of shooters B and C. Shooter C loads and shoots his own gun while shooter B is lazy and insists that A load his gun for him. The prisoner D will die (D = 1) if, and only if, either A loads B's gun (A = 1) and B shoots (B = 1) or C loads his gun and shoots (C = 1); that is, $D = (A \wedge B) \vee C$.

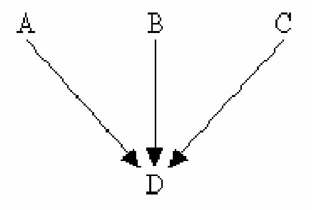

Figure 1: Firing Squad

For this example, if $(M, \vec{u}) \vDash (D = 1)$ then

$$\Delta(D = 1) = (A = 1 \wedge B = 1 \wedge C = 1)$$
$$\vee (A = 1 \wedge B = 1 \wedge C = 0) \vee (A = 0 \wedge B = 1 \wedge C = 1)$$
$$\vee (A = 1 \wedge B = 0 \wedge C = 1) \vee (A = 0 \wedge B = 0 \wedge C = 1) .$$

A term (conjunction of literals) that entails a sentence S is an *implicant* of S; an implicant that does not entail any other implicant is a *prime implicant*. The *prime implicant form* of a sentence is a disjunction of all its prime implicants and is unique. The prime implicant form of $\Delta(D = 1)$ is $\Delta(D = 1) = (A = 1 \wedge B = 1) \vee (C = 1)$.

With these preliminaries, Hopkins and Pearl (2003) prove the following theorem:

Theorem (prime implicant) In a causal model M with context $\mathcal{U} = \vec{u}$, let $X, Y \in \mathcal{V}$ with $X \in PA_Y$. If $(M, \vec{u}) \vDash (X = x \wedge Y = y)$ and the literal $X = x$ occurs in any prime implicant of $\Delta(Y = y)$ then the Halpern-Pearl definition of actual causation will classify $X = x$ as an actual cause of $Y = y$.

Note that the prime implicant theorem does *not* require that any other literals (if they exist) in any of the prime implicants of $\Delta(Y = y)$ to which $X = x$ belongs be satisfied (true) in $(M, \vec{u})$. For example, assume the context $\vec{u}$ in the firing squad example is such that C shoots and A loads B's gun, but B does not shoot. Since $(M, \vec{u}) \vDash (A = 1 \wedge D = 1)$ and $A = 1$ occurs in the prime implicant $(A = 1 \wedge B = 1)$ for $\Delta(D = 1)$, according to the prime implicant theorem, the Halpern-Pearl theorem should (counter-intuitively) classify A's loading of B's gun as a cause of D's death though B does not shoot. Indeed, taking $\vec{Z} = (A, D)$ and $\vec{W} = (B, C)$ and setting $\vec{W} = \vec{w}$ = (1,0)) satisfies conditions HC2(a) and (b) of the definition.

Hopkins and Pearl (2003) point out the similarity of the prime implicant form of a sentence with Mackie's INUS account of singular causation, according to which C is a cause of E iff C is an insufficient but non-redundant part of an unnecessary but sufficient condition for E (Mackie, 1974, pp. 61-62): "For instance, A loading B's gun is a necessary part of a sufficient condition to ensure the prisoner's death. In terms of the prime implicant logical form, sufficient conditions map to implicants. For instance, $A = 1 \wedge B = 1$ is a sufficient condition for $D = 1$. Furthermore, since $A = 1 \wedge B = 1$ is a *prime* implicate (hence no subset of its conjuncts is an implicate), we observe that both $A = 1$ and $B = 1$ are necessary parts of this sufficient condition. Hence any atomic expression that appears in a prime implicate satisfies the INUS condition."

Accepting the mapping of sufficient conditions to implicants, then Mackie's INUS test (Kim, 1993) requires that for $A = 1$ to be a cause of $D = 1$, not only must $A = 1$ occur as an atomic proposition (or literal) in some prime implicant for $D = 1$ (i.e., be an INUS condition for $D = 1$) but also that every other atomic proposition in that implicant be satisfied. This part of Mackie's analysis is consistent with the NESS test (see Baldwin, 2003). It follows then that the Halpern-Pearl definition is less restrictive than both Mackie's INUS analysis and Wright's NESS test. As Hopkins and Pearl say, their prime implicant theorem exposes that the Halpern-Pearl definition is over permissive. It is at least too permissive to formally capture the meaning of the NESS test.

Preliminaries to the New Definition

We take as a consequence of Hopkin's and Pearl's prime implicant theorem that if a structural definition of actual causation employing the counterfactual strategy is to capture the meaning of the NESS test, the choice of which variables in the model may have their values fixed must be controlled by the relationship of belonging to the same minimal sufficient condition (set of conditions) for an effect. Pearl (2000) suggests that this information may already be encoded in the structural language. In discussing Mackie's (1974) scenario of the ill-fated desert traveler—where a traveler has two enemies, one who poisons (p = 1; variables are chosen to be consistent with Pearl's exposition) the traveler's water canteen and the other, unaware of the poisoning, shoots and empties the traveler's canteen (x = 1) as a result of which the traveler dies—Pearl (2000, p. 312) considers the structural equations $y = x \vee px'$ (x' represents $\neg x$) and the *logically* equivalent $y = x \vee p$ and states, "Here we see in vivid symbols the role played by structural information. Although it is true that $x \vee x'p$ is logically equivalent to $x \vee p$, the two are not structurally equivalent; $x \vee p$ is completely symmetric relative to exchanging x and p, whereas $x \vee x'p$ tells us that, when x is true, p has no effect whatsoever—not only on y, but also on any of the intermediate conditions that could potentially affect y. It is this asymmetry that makes us proclaim x and not p to be the cause of death."

Hausman and Woodward (1999) shed more light on the relationship between the structural information that the causal relation between an independent variable and dependent variable in a structural equation depends (or does not depend) on one or more other independent variables in the equation (e.g., the difference between $x \vee p$ and $x \vee x'p$) and minimal sets of sufficient conditions. They point out that a structural equation may capture more than one causal mechanism; terms in additive relations in a single structural equation may represent distinct causal mechanisms. For example, x and $x'p$

represent distinct causal processes ending in the traveller's death. Hausman and Woodward (1999, p. 547) say that a system of structural equations with additive terms where each *term* in each equation represents a distinct causal mechanism ("and thus acts independently of the mechanisms registered by other terms") exhibits *coefficient invariance*, which is violated when the terms in the structural equations are not additive or when the causal relationship between two variables depends on the level of a third. Hausman and Woodward (1999, p. 547) then say, "If one thinks of individual causes of some effect Y as conjuncts in a minimal sufficient condition for Y (or the quantitative analogue thereof)—that is, as 'conjunctive causes'—then the relationship between an effect and its individual causes will not satisfy coefficient invariance"

Hausman and Woodward define coefficient invariance as a global property of systems of structural equations and as a means of identifying distinguishable (if not distinct in the sense of shared variables) mechanisms within a single structural equation—which we require—their explication of the concept is too strict. (Hausman and Woodward (1999, p. 547) say that coefficient invariance holds only when *every* variable in a structural equation belongs to a single additive term ruling out, for example, equations of the form $y = x \vee px'$.) For our purposes, it is enough that for a set of additive structural equations, expressed in sum of products form, each term in an equation represents a separate mechanism, a separate set of minimal sufficient conditions ("or the quantitative analogue thereof"), for the effect represented by the equation's dependent variable. We call this property *term modularity*. A causal model whose structural equations satisfy this property allows for the development of a criterion, ultimately derivative of Hausman and Woodward's concept of coefficient invariance, for determining what variables should be held fixed and what variables may be altered in testing for counterfactual dependence between an effect variable and one of its (putative) causal variables in the model. This, in turn, will allow for a new structural definition of actual causation that avoids the problem identified by Hopkins and Pearl and that formalizes the NESS test in the context of a scenario modelled by a causal world in the structural language.

The "term modularity" criterion encodes the relationship, among the parents of a variable, of being components of distinct component causal mechanisms (or elements of minimal sufficient conditions for the variable). For distinct variables *X*, *Y*, and *Z*, where *X* and *Y* occur as independent variables in the structural equation (parents) for Z in a causal world $(M,\vec{u})$, X *is coefficient invariant to* Y *for term* T if $(M,\vec{u}) \models \neg(T = 0)$ (i.e., the term is satisfied, or non-zero in quantitative contexts, in $(M,\vec{u})$), *X* occurs as a literal in *T*, and *Y* is not a variable in *T* (symbolically, $coin_{\vec{u}}^{T}(Z;X \mid Y)$; note that *X* is a literal of the form $X = x$ where $(M,\vec{u}) \models (X = x)$ while *Y* is a variable). When $coin_{\vec{u}}^{T}(Z;X \mid Y)$ the causal relation between X and Z does not depend on the value of Y in context $\vec{u}$. Locally, to avoid satisfying actually unsatisfied terms (minimal sufficient sets) as happens with the Halpern-Pearl definition, between a variable X and its parents ($Y_1,\ldots,Y_n$), in testing whether $Y_i = y_i$ is an actual cause of $X = x$, $Y_i = y_i$ should belong to a satisfied term *T* and only parent variables Y_j that Y_i is coefficient invariant to for *T* in the equation for X ($coin_{\vec{u}}^{T}(X;Y_i \mid Y_j)$) should be allowed to have their values altered.

Returning to Hopkins and Pearl's firing squad example, where the sole structural equation is $D = (A \wedge B) \vee C$, $(M,\vec{u}) \models (A \wedge B)$ and it is not the case that *A* is coefficient invariant to *B* for $T = (A \wedge B)$ in the equation for *D* ($\neg coin_{\vec{u}}^{(A\wedge B)}(D;A \mid B)$). Therefore, in the context such that *A* = *C* = 1 and *B* = 0, to test whether *D* = 1 is counterfactually dependent on *A* = 1, the value of *B* may not be altered. Since $(A \wedge B)$ is the only satisfied term in the equation for *D* in which *A* (i.e., $A = 1$) occurs, contrary to the Halpern-Pearl approach, it is not possible to modify the model so that *D* is counterfactually dependent on *A* in a scenario where *B* = 0.

Globally in a causal model with context $\mathcal{U} = \vec{u}$, when distinct variables Y_1 and Y_2 occur as common parents of distinct variables X_1 and X_2 it can happen that there exists a term T_{X_1} in the equation for X_1 such that $coin_{\vec{u}}^{T_{X_1}}(X_1;Y_1 \mid Y_2)$ but any satisfied term T_{X_2} in the equation for X_2 that includes Y_1 also includes Y_2 ($\neg coin_{\vec{u}}^{T_{X_2}}(X_2;Y_1 \mid Y_2)$); that is, Y_1 is coefficient invariant to Y_2 for some term in the equation for X_1 but not in the equation for X_2. In that case, if Y_1 is part of the "active causal process" being tested, before allowing the value of Y_2 to be altered, it is necessary to interfere directly in the equation for X_2 by substituting a constant y_2 for Y_2 in the equation for X_2 where $Y_2 = y_2$ in the unaltered model (i.e., fix Y_2 at its actual value, $(M,\vec{u}) \models (Y_2 = y_2)$). This avoids the possibility of the counterfactual or original values of Y_1 interacting with non-actual values to satisfy non-actually satisfied minimal sufficient sets for some variable, the problem that plagues the Halpern-Pearl definition. This process must be repeated for all X_i where for all satisfied terms T_{X_i} including $\neg coin_{\vec{u}}^{T_{X_i}}(X_i;Y_1 \mid Y_2)$. Only then should altering the value of Y_2 be allowed.

New Structural Definition of Actual Causation

A *causal route* $\vec{R} = \langle C, D_1, \ldots, D_n, E \rangle$ between two variables C and E in $\mathcal{V}$ is an ordered sequence of variables such that each variable in the sequence is in $\mathcal{V}$ and a parent of its successor in the sequence.

For a causal mode M with route $\vec{R} = \langle C, D_1, \ldots, D_n, E \rangle$ and a sequence of terms $\vec{T} = \langle T_{D_1}, \ldots, T_{D_n}, T_E \rangle$, where T_X is a satisfied term in the equation for X, the *submodel relative to* $\vec{R}$ *and* $\vec{T}$ *in context* $\vec{u}$ (denoted $M^{\vec{T}}_{[\vec{R},\vec{u}]}$) is derived from $(M, \vec{u})$ as follows: for distinct $X, Y, W \in \mathcal{V}$ with $X \in \vec{R} - E$, $Y \notin \vec{R}$, and $W \neq C$, if $\neg coin^{T_W}_{\vec{u}}(W; X \mid Y)$ replace the function F_W for by the function that results when Y is replaced with a constant y where $(M, \vec{u}) \vDash (Y = y)$.

Definition (*actual cause*; *new version*) $C = c$ is an actual cause of $E = e$ in $(M, \vec{u})$ if the following conditions hold:
AC1. $(M, \vec{u}) \vDash (C = c \wedge E = e)$
AC2. There exists a route $\vec{R} = \langle C, D_1, \ldots, D_n, E \rangle$ in M, a sequence of satisfied terms $\vec{T} = \langle T_{D_1}, \ldots, T_{D_n}, T_E \rangle$, and a setting $\vec{w}$ for $\vec{W} = \mathcal{V} - \vec{R}$ and a setting $c' \neq c$ for C such that:

$(M^{\vec{T}}_{[\vec{R},\vec{u}]}, \vec{u}) \vDash [C \leftarrow c', \vec{W} \leftarrow \vec{w}] \neg (E = e)$, and

$(M^{\vec{T}}_{[\vec{R},\vec{u}]}, \vec{u}) \vDash [C \leftarrow c, \vec{W} \leftarrow \vec{w}](E = e)$.

Because there are no causal interaction effects between variables in $\vec{R}$ and $\vec{W}$ in $M^{\vec{T}}_{[\vec{R},\vec{u}]}$, by the construction of $M^{\vec{T}}_{[\vec{R},\vec{u}]}$ (variables in $\vec{R}$ are coefficient invariant to all variables in $\vec{W}$ by definition of $M^{\vec{T}}_{[\vec{R},\vec{u}]}$), the setting $\vec{W} \leftarrow \vec{w}$ cannot "contaminate" the test of counterfactual dependence in AC2 in the sense of satisfying a non-actually satisfied minimal sufficient set of conditions.

In practice, it rarely happens that a literal X occurs in more than one satisfied term in a structural equation; a non-quantitative equation having more than one satisfied term with distinct literals only occurs itself in cases of duplicative causation. To avoid the cumbersome and somewhat confusing terminology, subsequently, unless the context requires otherwise (as in the analysis of the pollution cases below), the choice of the sequence $\vec{T}$ will be left as implied by the analysis of the scenario and the superscript $\vec{T}$ left out of the notations $coin^{T}_{\vec{u}}(Z; X \mid Y)$ and $M^{\vec{T}}_{[\vec{R},\vec{u}]}$.

Suppose that C is an actual cause of E in $(M, \vec{u})$. Then there exists a route $\vec{R} = \langle C, D_1, \ldots, D_n, E \rangle$ and a sequence of satisfied terms $\vec{T} = \langle T_{D_1}, \ldots, T_{D_n}, T_E \rangle$ satisfying condition AC2 of the new definition of actual causation. The *active causal process relative to* $\vec{R}$ *and* $\vec{T}$ *in* $\vec{u}$ (denoted $ACP^{\vec{T}}_{[\vec{R},\vec{u}]}$) is the set $\vec{R} \cup \{X_i\}$ where $X_i \in \mathcal{V} - \vec{R}$, $Y \in \vec{R} - E$, $Z \in \vec{R} - A$, and $\neg coin^{T_Z}_{\vec{u}}(Z; Y \mid X_i)$. That is to say, $ACP^{\vec{T}}_{[\vec{R},\vec{u}]}$ is the subset of variables in $\mathcal{V}$ that have their valued fixed in forming $M^{\vec{T}}_{[\vec{R},\vec{u}]}$ that are parents of a variable in $\vec{R}$. (Again, the term specific terminology and the accompanying superscripts will be discarded where the context does not require them.)

Examples of NESS and the New Definition

Preemptive causation

Wright (1985, p 1795) considers two scenarios: in the first, *D* shoots and kills *P* before *P* drinks tea fatally poisoned by *C* and, in the second, *D* shoots and instantly kills *P* after *P* drinks tea fatally poisoned by *C* but before the poison takes effect. In the first scenario, in Wright's (1985, p 1795) NESS analysis, *D*'s shot was necessary for the sufficiency of a set of actual antecedent conditions that did not include the poisoned tea. Conversely, *C*'s poisoning of the tea was not a necessary element of any sufficient set of actual antecedent conditions. A set that included the poisoned tea but not the shooting would be sufficient only if *P* actually drank the tea, but this was not an actual condition. The shooting preempted the potential causal effect of the poisoned tea.

In this scenario, the story of death by poisoning would have occurred (the intake of the poison through consumption of the tea will have occurred) but for *D* shooting *P*. This is reflected in the following causal model. The model has the following propositional variables: *DS* represents "*D* shoots," *PT* represents "*C* poisons the tea," *CP* represents "*P* consumes poison," and *PD* for "*P* dies." The structural equations are $CP = \neg DS \wedge PT$ and $PD = DS \vee CP$. The causal diagram corresponding to these equations is represented in Figure 2.

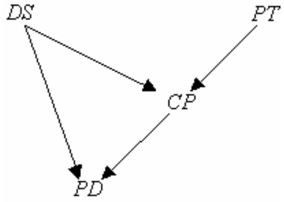

Figure 2: Causal diagram for the poisoned tea scenario

To show that $DS = 1$ is an actual cause of $PD = 1$, let $\vec{R} = \langle DS, PD \rangle$ for condition AC2. Since $coin_{\vec{u}}(PD; DS \mid CP)$, $M_{[\vec{R},\vec{u}]} = (M, \vec{u})$ and therefore $\vec{W} = (PT, CP)$. Setting $\vec{W} = (0,0)$ then satisfies conditions AC2(a) and (b). Note that $ACP_{[\vec{R},\vec{u}]} = \vec{R}$ and the NESS set including $DS = 1$ for $PD = 1$ in $(M, \vec{u})$ is just $\{DS = 1\}$, as it is with Wright's analysis.

Suppose that the context was such that *D* does not shoot ($\neg DS$) but P still poisons the tea. Then $CP = 1$ and $PD = 1$ and it is straightforward to show that $PT = 1$ is a cause of $PD = 1$ by letting $\vec{R} = \langle PT, CP, PD \rangle$ in condition AC2. Note, however, that since $\neg coin_{\vec{u}}(CP; PT \mid DS)$, $ACP_{[\vec{R},\vec{u}]} = \{PT, CP, DS, PD\}$ and the NESS set including $PT = 1$ for $PD = 1$ in $(M, \vec{u})$ is $\{PT = 1, CP = 1, DS = 0\}$: the absence of the preempting condition *DS* must be included.

For the second example, Wright's (1985, p 1795) NESS analysis of why *D*'s shooting was a cause of *P*'s death is the same as that for the first example; as to whether *C*'s poisoning of the tea was a cause: "Even if P actually had drunk the poisoned tea, C's poisoning of the tea still would not be a cause of P's death if the poison did not work instantaneously but the shot did. The poisoned tea would be a cause of P's death only if P drank the tea and was alive when the poison took effect. That is, a set of actual antecedent conditions sufficient to cause P's death must include poisoning of the tea, P's drinking the poisoned tea, and P's being alive when the poison takes effect. Although the first two conditions actually existed, the third did not. D's shooting P prevented it from occurring. Thus, there is no sufficient set of actual antecedent conditions that includes C's poisoning of the tea as a necessary element. Consequently, C's poisoning of the tea fails the NESS test. It did not contribute to P's death."

A causal model for this scenario differs from the previous one by the addition of a variable *PTE* for "the poison takes effect." The structural equation for *PD* becomes $PD = DS \vee PTE$ and the equation for *CP* becomes $CP = PT$. As with Wright's NESS analysis, the proof that $DS = 1$ is an actual cause of $PD = 1$ would be essentially the same as with the previous example.

Duplicative Causation Scenarios

Among the duplicative causation cases, of particular interest are a group of pollution cases where defendants were found liable though none of their individual acts (their "contributions" to the pollution) was sufficient, or necessary given the contributions of the other defendants, to produce the plaintiff's injuries (some adverse effect on the use of his property).[1] Wright (1985, p 1793) applies the NESS test to an idealized example in which, five units of pollution are necessary and sufficient for the plaintiff's injury and seven defendants discharge one unit each. The NESS test requires only that a defendant's discharge be necessary for the sufficiency of *a* set of actual antecedent conditions, and that (Wright 1985, p 1795) "each defendant's one unit was necessary for the sufficiency of a set of actual antecedent conditions that included only four of the other units, and the sufficiency of this particular set of actual antecedent conditions was not affected by the existence of two additional duplicative units."

In fact, in this sense, for each defendant's discharge there are fifteen distinct actually sufficient sets of antecedent conditions, one for each possible choice of any four of the 6 remaining defendant's units of pollution.

The causal model for this example has variables X_i, $i = 1, \ldots, 7$ $i = 1, \ldots, 7$, representing whether defendant *i* contributed his one unit of pollution $(X_i = 1)$ or not $(X_i = 0)$. The single structural equation is

$$DP = (X_1 = 1 \wedge X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1 \wedge X_5 = 1) \vee \ldots \vee \\ (X_7 = 1 \wedge X_6 = 1 \wedge X_5 = 1 \wedge X_4 = 1 \wedge X_3 = 1) .$$

It consists of 21 terms where each term is a conjunction of 5 of the 7 literals $X_i = 1$. Since each literal $X_i = 1$ is satisfied in the given scenario $(M, \vec{u})$, each literal occurs in 15 satisfied terms in conjunction with 4 of the remaining 6 X_i or, equivalently, each literal $X_i = 1$ occurs in 15 terms without conjuncts involving 2 of the remaining 6 variables. Thus, for any $X_i = 1$ and variables X_k, X_l $(i \neq k \neq l)$, there exists some term T_{DP} with $coin_{\vec{u}}^{T_{DP}}(DP; X_i \mid X_k, X_l)$.

Without loss of generality, to show that each defendant's pollution discharge is an actual cause of $DP = 1$, let $i = 1$ and choose T_{DP} so that $coin_{\vec{u}}^{T_{DP}}(DP; X_1 \mid X_6, X_7)$ (i.e.,

[1] For example, Wright (2001, p 1100) cites the case of Warren v. Parkhurst, 92 N.Y.S. 725 (N.Y. Sup. Ct. 1904), aff'd, 93 N.Y.S. 1009 (A.D.1905), aff'd, 78 N.E. 579 (N.Y. 1906), where each of twenty-six defendants discharged "nominal" amounts of sewage into a creek which individually were not sufficient to destroy the use of downstream plaintiff's property but the stench of the combined discharges was sufficient.

$T_{DP} = (X_1 = 1 \wedge X_2 = 1 \wedge X_3 = 1 \wedge X_4 = 1 \wedge X_5 = 1)$). Then with $\vec{R} = \langle X_1, DP \rangle$ and $\vec{T} = \langle DP \rangle$, the equation for DP in $M^{\vec{T}}_{[\vec{R},\vec{u}]}$ is

$$DP = (X_1 = 1) \vee (X_6 = 1) \vee (X_7 = 1) \vee (X_1 = 1 \wedge X_6 = 1) \vee (X_1 = 1 \wedge X_7 = 1)$$
$$\vee (X_6 = 1 \wedge X_7 = 1) \vee (X_1 = 1 \wedge X_6 = 1 \wedge X_7 = 1)$$

Since, in $M^{\vec{T}}_{[\vec{R},\vec{u}]}$, DP is a trivial function of the variables in $\{X_2, \ldots, X_5\}$, for $\vec{W} = \mathcal{V} - \vec{R} = \{X_i\}$, $i = 2, \ldots, 7$, of condition AC2 of the new definition, only the settings for X_6 and X_7 matter. Setting $(X_6, X_7) = (0,0)$, $X_1 = 1$ is easily seen to satisfy the counterfactual test of condition AC2.

Note that $ACP^{\vec{T}_{DP}}_{[\vec{R},\vec{u}]} = \{X_1, X_2, \ldots, X_5, DP\}$ and, in the causal world $(M, \vec{u})$, $X_1 = 1$ is necessary for the sufficiency of the set including defendant one's discharge ($X_1 = 1$) and only four other discharges.

Double Omission Cases

A class of cases that have proved problematic for the NESS test, the so-called double omission cases, suggest that modeling is an important aspect of a NESS enquiry in practice: "Some of the most difficult overdetermined-causation cases, at least conceptually, are those involving multiple omissions, which usually involve failures to attempt to use missing or defective safety devices or failures to attempt to read or heed missing or defective instructions or warnings." (Wright 2001, pp. 1123-1124).

Wright (1985, p. 1801; 2001, p. 1124 ff.) considers in detail the case of *Saunders System Birmingham Co. v. Adams*[1] where a car rental company negligently failed to discover or repair bad brakes before renting a car out. The driver who rented the car then negligently failed to apply the brakes and struck a pedestrian. In general, courts have held that individuals who negligently fail to repair a device (or provide proper safeguards or warnings) are not responsible when (negligently) no attempt was made to use the device (or use the safeguards or observe the warnings). According to Wright (2001, p. 1124), the court's decisions reflect a "tacit understanding of empirical causation in such situations": not providing or repairing a device (or not providing proper safeguards or warnings) can have no causal effect when no attempt was or would have been made to use the device (or use the safeguard or observe the warning)—unless no attempt was made because it was known that the device was inoperative (or the safeguards or warnings were inadequate).

Wright's (1985, p. 1801) NESS analysis (where *D* represents the driver, *C* represents the car rental company, and *P* represents the pedestrian) is as follows: *D*'s negligence was a preemptive cause of *P*'s injury. *C*'s negligence did not contribute to the injury. *D*'s not applying the brakes was necessary for the sufficiency of a set of actual antecedent conditions not including *C*'s failure to repair the brakes. The sufficiency of this set was not affected by *C*'s failure to repair the brakes: "A failure to try to use brakes will have a negative causal effect whether or not the brakes are defective." *C*'s failure to repair the brakes was not a necessary element of any set of antecedent actual conditions that was sufficient for the occurrence of the injury: "Defective brakes will have an actual causal effect only if someone tries to use them." The effect of *C*'s failure to repair the brakes was preempted by *D*'s failure to try to use them.

Notice that interchanging *C* and *D*'s negligent acts in this argument results in an apparently equally plausible argument for *C*'s negligence being a preemptive cause of *P*'s injury. According to Wright (2001, p.1125): "At the time that I wrote this explanation, I was aware that it was too brief and cryptic, relied upon an insufficiently elaborated notion of causal sufficiency and 'negative causal effect,' and therefore could seemingly be reversed to support the opposite causal conclusions merely by switching the references to the two omissions. Nevertheless, I thought it roughly stated the correct analysis in very abbreviated form."

Binary variables for this causal model might be: *RB* for "repairs brakes", *AB* for "applies brakes", *BO* for "brakes operate", and *HP* for "pedestrian is hit". The question is, what are the structural equations? The structural equations suggested by the argument that in the NESS analysis the roles of *C* an *D* are symmetrical might be $\neg BO = \neg RB \vee \neg AB$ and $PH = \neg BO$.

It is easy to see that the new definition will classify both $\neg RB$ and $\neg AB$ as actual causes of *PH* for this model. On the other hand, suppose the structural equations are $BO = RB \wedge AB$ and $PH = \neg BO$. In that case the new definition will classify neither *RB* nor *AB* as a cause of *PH*. This model captures the intuition that not repairing the brakes is not a cause of the pedestrian being hit if the brakes are not applied but also suggests that not applying the brakes cannot cause the striking of the pedestrian if the brakes are not operative. Notice, however, that in Wright's analysis there is the suggestion of a mechanism that neither of these models includes: not using the brakes will have a causal effect whether or not the brakes are not repaired. In other words, there are two distinct mechanisms for the pedestrian being hit; confusion arises because not braking just happens to play a part in both. On this latter analysis, $BO = RB \wedge AB$ and $PH = \neg BO \vee \neg AB$ are the model equations. The causal diagram for this model is represented in Figure 3.

[1] Saunders Sys. Birmingham Co. v. Adams, 117 So. 72 (Ala. 1928).

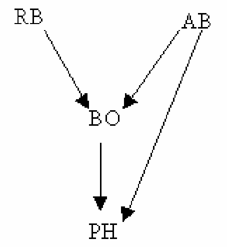

Figure 3: Causal diagram for the braking scenario

Indeed, for this model the new definition will classify $\neg AB$ as a cause of *PH* but not $\neg RB$. It is this missing mechanism that lies behind the intuitive and analytic confusion in the double omission cases. Wright's initial NESS argument was not incorrect but only lacked an adequate language to represent the causal dynamics of the scenario.

Conclusions

Developing a definition of actual or token causation conforming to intuitive judgments is an active problem for both legal scholarship and AI. As an element in the determination of legal responsibility, courts have been required to develop a practical test for actual causation. The accepted test, the 'but-for' test, is limited in its application. Wright's NESS test appears to successfully address these limitations. The NESS test itself requires a counterfactual test. The language of structural models allows for the formal representation of counterfactual arguments. We have presented a formal definition of actual causation in the language of structural models, which we believe captures the essential meaning of the NESS test while successfully avoiding the weaknesses inherent an earlier structural definition of Halpern and Pearl.

Acknowledgements

This research was supported by the second author's NSERC Discovery Grant and a University of Saskatchewan Graduate Scholarship. This version corrects errors appearing in the CSCSI-2004 version.

References

Ashley, K. D. 1990. Modeling legal arguments: reasoning with cases and hypotheticals. Cambridge : MIT Press, 329 pages

Baldwin, Richard 2003. A Structural Model Interpretation of Wright's NESS test. Department of Computer Science, University of Saskatchewan, MSc thesis [Online]. Available: http://library.usask.ca/theses/available/etd-09152003-154313/ [Accessed 2003, Nov. 28].

Baldwin, R., and Neufeld, E. 2003. On the Structure Model Interpretation of Wright's NESS Test. In *Proceedings of AI 2003,* Halifax (June 2003) LNAI 2671 9-23

Wright, R.W. 1985. Causation in tort law. California Law Review, 73, pp. 1735-1828.

Halpern, J.Y., and Pearl, J. 2000. Causes and explanations: a structural-model approach. Retrieved September 3, 2001 from http://www.cs.cornell.edu/home/halpern/topics.html#rau (Part I, Causes, appears in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, San Francisco, CA: Morgan Kaufmann, 194-202, 2001.)

Hart, H.L.A., and Honoré, A.M. 1985. *Causation in the law* (2nd ed.). Oxford University Press.

Hopkins, M., and Pearl, J. 2003. Clarifying the Usage of Structural Models for Commonsense Causal Reasoning. In Proceedings of the 2003 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford University, March 24-26, 2003.

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

Pearl, J. 1998. On the definition of actual cause. Technical Report (no. R-259), Department of Computer Science, University of California, Los Angeles.

Pearl, J. 2000. *Causality: Models, reasoning, and inference*. New York: Cambridge University Press.

Wright, R.W. 1985. Causation in tort law. *California Law Review*, 73, pp. 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, pp. 1001-1077.

Wright, R.W. 2001. Once more into the bramble bush: Duty, causal contribution, and the extent of legal responsibility [Electronic version]. *Vanderbilt Law Review*, 54 (3), pp. 1071-1132.