DIFFERENCE-MAKING CAUSATION*

Lewis thinks of causes as difference makers. Whether or not a cause occurs makes a difference as to whether or not its effect occurs. He thus aimed to analyze causation in terms of counterfactual dependence.[1] An event *e* counterfactually depends on another event *c* if and only if (iff), had *c* not occurred, *e* would not have occurred. On Lewis's analysis, an occurring event *c* is a cause of a distinct occurring event *e* if *e* counterfactually depends on *c*. Among the accounts in the tradition of Lewis, counterfactual dependence between distinct occurring events is taken to be sufficient for causation.[2] The strategy of counterfactual accounts is thus to ask "what would happen if the putative cause had been absent?" Under this counterfactual assumption they claim causation if the presumed effect is absent as well. An event is a cause in virtue of making this difference.

In this paper, we put forth a counterfactual analysis of causation. Here is the gist: an event *c* is a cause of another event *e* just in case both events occur, and—after taking out the information whether or not *c* and *e* occur—*e* would not occur if *c* were not to occur. We will show that the analysis successfully captures a wide range of causal scenarios, including switches, preemption, and two scenarios of double prevention. To date, there is no other counterfactual account that can solve this set of scenarios.

*We would like to thank Katie Steele, Atoosa Kasirzadeh, Cei Maslen, Alan Hájek, Phil Dowe, Daniel Stoljar, and an anonymous reviewer for this JOURNAL for helpful comments. We are grateful to audiences of the Philosophy Departmental Seminar at The Australian National University, of the 2019 Annual Conference of the New Zealand Association of Philosophers, and of the conference "Bayesian Epistemology: Perspectives and Challenges" at the Munich Center for Mathematical Philosophy. The work has been supported by the Humanising Machine Intelligence project. The authors contributed equally to this paper.

[1] David Lewis, "Causation," this JOURNAL, LXX, 17 (Oct. 11, 1973): 556–67.

[2] See, for instance, Lewis's "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000): 182–97; M. Ramachandran, "A Counterfactual Analysis of Causation," *Mind*, CVI, 422 (April 1997): 263–77; Christopher Hitchcock, "The Intransitivity of Causation Revealed in Equations and Graphs," this JOURNAL, XCVIII, 6 (June 2001): 273–99; Stephen Yablo, "De Facto Dependence," this JOURNAL, XCIX, 3 (March 2002): 130–48; James Woodward, *Making Things Happen: A Theory of Causal Explanation* (Oxford: Oxford University Press, 2003); Ned Hall, "Two Concepts of Causation," in John Collins, Ned Hall, and L. A. Paul, eds., *Causation and Counterfactuals* (Cambridge, MA: MIT Press, 2004), pp. 225–76; Ned Hall, "Structural Equations and Causation," *Philosophical Studies*, CXXXII (2007): 109–36; and Joseph Y. Halpern and Judea Pearl, "Causes and Explanations: A Structural-Model Approach. Part I: Causes," *British Journal for the Philosophy of Science*, LVI (2005): 843–87.

0022-362X/21/1812/680–701

For extant counterfactual accounts, switching scenarios mean trouble. Consider the scenario depicted in Figure 1.

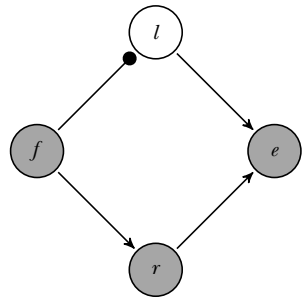


Figure 1.

In this simple switch, neuron f fires. This excites r's firing, which in turn excites neuron e to fire. At the same time, f's firing inhibits the excitation of l, which would have been excited in case f had not fired. e fires if either one of l or r fires. In brief, f determines which one of l and r is firing, and thus acts like a switch. f's firing makes no difference as to whether or not e fires. Hence, f should not count as a cause of e's firing—in particular on a counterfactual account of causation.

Surprisingly, Lewis's analysis misclassifies f as a cause of e. He says causation is the transitive closure of non-backtracking counterfactual dependence between occurring events. A backtracking counterfactual retraces some past causes from an effect. Now, f, r, and e occur, and both r counterfactually depends on f in a non-backtracking way and e does so on r. Barring backtracking, if r had not fired, e would not have fired. By the transitive closure imposed on the one-step causal dependences, Lewis is forced to say that f is a cause of e.[3]

Progress has been made. More recent counterfactual accounts of causation have used causal models to capture scenarios that defy Lewis's analysis. A causal model $\langle M, V\rangle$ represents a causal scenario by a set of structural equations M and a variable assignment V. For the above scenario, M is given by the structural equations $\mathrm{l} = \neg\mathrm{f}$, $\mathrm{r} = \mathrm{f}$, and $\mathrm{e} = \mathrm{l} \vee \mathrm{r}$. l fires iff f does not; r fires iff f does; and e fires iff l or r does. The variable assignment V may be given by the set $\{f, \neg l, r, e\}$, which indicates that all neurons except l fire. In this

[3] Lewis still imposes transitivity on his refined analysis of causation in "Causation as Influence," this JOURNAL, XCVII, 4 (April 2000): 182–97. As a result, the refined analysis is also forced to say that f is a cause of e in the switching scenario.

causal model, we may set the variable f to $\neg f$ and propagate forward the changes effected by this intervention. Given that $\neg f$, the structural equations determine that e. Even if f had not fired, e would have fired nonetheless. Hence, f does not count as a cause of e on counterfactual accounts, or so it seems.

Most of the extant counterfactual accounts share a strategy. They test for counterfactual dependence while keeping certain events or variable assignments fixed. Hitchcock, for example, proposes that c is a cause of e relative to a causal model iff there is an active causal route from c to e in the causal model.[4] Now, there is an active route from f over r to e, and keeping the off-path variable l fixed at its actual value induces a counterfactual dependence of e on f. So Hitchcock's account, like many other accounts, still fails to solve the switch scenario.[5]

We approach the switch scenario in a different way. Our analysis centers on the notion of a causal model that provides only partial information as to which events occur, but complete information about the dependences between the events. To outline the analysis: c is a cause of e relative to a causal model $\langle M, V\rangle$ iff

1. c and e are true in $\langle M, V\rangle$, and
2. there is $V' \subset V$ such that $\langle M, V'\rangle$ contains no information about c and e, but in which $\neg e$ would be true if $\neg c$ were.

By these conditions, we test whether e counterfactually depends on c in a causal model that is uninformative on c and uninformative on e. Causation is thus analyzed as uninformative counterfactual dependence.

Why is f's firing no cause of e's firing in the switch scenario? Well, there is simply no causal model $\langle M, V'\rangle$ that contains no information about e. Even if there is no information at all as to which events occur, the information about the dependences between the events is sufficient for e's occurrence. No matter whether f or $\neg f$ is actual, e occurs according to the structural equations. Unlike other counterfactual accounts, our analysis thus captures the difference-making judgment about the switch scenario.

[4] See Hitchcock, "The Intransitivity of Causation Revealed in Equations and Graphs," *op. cit.*

[5] In fact, all of the counterfactual accounts cited in footnote 2 misclassify f as a cause of e, except Hall's account in "Structural Equations and Causation," *op. cit.* However, that account is inadequate for different reasons, as Hitchcock shows in "Structural Equations and Causation: Six Counterexamples," *Philosophical Studies*, CXLIV (2009): 391–401. Halpern's modification of the Halpern and Pearl account in "Causes and Explanations," *op. cit.*, still cannot solve the switch scenario; see his *Actual Causality* (Cambridge, MA: MIT Press, 2016), p. 25.

We have solved the switch scenario without invoking a transitive closure of counterfactual dependences. Likewise, we have no need to keep certain variable assignments fixed by intervention when testing for counterfactual dependence. In fact, this strategy, common to many counterfactual accounts employing causal models, is at fault for the misclassification of f as a cause of e in the switch scenario. By contrast, we merely require that there must be a causal model uninformative on the cause and the effect, in which the effect counterfactually depends on the cause.

It remains to show that our counterfactual analysis of causation gives the intuitively correct results for various other causal scenarios. We thus refine our analysis and apply it to causal scenarios in what follows. In section I, we outline our account of causal models and state a preliminary version of our analysis. In section II, we make the relation between neuron diagrams and our causal models explicit before we tackle a number of causal scenarios. In section III, we deal with symmetric overdetermination and state our final analysis.

I. A CAUSAL MODEL ANALYSIS OF CAUSATION

Our causal models have two components: a set M of structural equations and a consistent set V of literals. Where p is a propositional variable, p is a positive literal and $\neg p$ a negative literal. p and $\neg p$ represent the values p can take. We give literals thus a semantic role.

The literals in V denote which events occur and which do not, that is, which events and absences are actual. $p \in V$ means that the event corresponding to p occurs. $\neg p \in V$, by contrast, means that no token event p of the relevant type occurs. Since the set of literals is consistent, it cannot be that both p and $\neg p$ are in V. Arguably, an event cannot both occur and not occur at the same time.

A structural equation denotes whether an event would occur if some other events were or were not to occur. Where p is a propositional variable and ϕ a propositional formula, we say that the following is a structural equation:

$$\mathrm{p} = \phi.$$

Each logical symbol of ϕ is either a negation, a disjunction, or a conjunction. ϕ can be seen as a truth function whose arguments represent occurrences and non-occurrences of events. The truth value of ϕ determines whether p or $\neg p$.

In Figure 1, there are arrows from the neurons l and r to the neuron e. The arrows represent that the value of the propositional variable e is determined by the values of the propositional variables l and r. The specific structural equation is $\mathrm{e} = \mathrm{l} \vee \mathrm{r}$. This equation encodes

four conditionals: (i) if l and r were actual, e would be; (ii) if l and $\neg r$ were actual, e would be; (iii) if $\neg l$ and r were actual, e would be; and finally, (iv) if $\neg l$ and $\neg r$ were actual, $\neg e$ would be. Structural equations thus describe dependences between actual and possible token events. For readability, we will represent causal models in two-layered boxes. The causal model of the switch scenario, for example, is given by $\langle\{\mathrm{l} = \neg\mathrm{f}, \mathrm{r} = \mathrm{f}, \mathrm{e} = \mathrm{l} \vee \mathrm{r}\}, \{f, \neg l, r, e\}\rangle$. We will depict such causal models $\langle M, V\rangle$ in a box, where the upper layer shows the set M of structural equations and the lower layer the set V of actual literals. For the switch scenario, we obtain:

$\mathrm{l} = \neg\mathrm{f}$
$\mathrm{r} = \mathrm{f}$
$\mathrm{e} = \mathrm{l} \vee \mathrm{r}$
$f, \neg l, r, e$

We say that a set V of literals satisfies a structural equation $\mathrm{p} = \phi$ just in case both sides of the equation have the same truth value when plugging in the literals in V. In more formal terms:

Definition 1. V satisfies ϕ and V satisfies $\mathrm{p} = \phi$
Let V be a set of literals that completely specifies which events in the causal scenario occur. Let v be the truth value assignment to the propositional variables (of the causal model) that satisfies V. We say that V satisfies a propositional formula ϕ iff ϕ is true on v. We say that V satisfies a structural equation $\mathrm{p} = \phi$ iff p and ϕ have the same truth value on v.

In the switch scenario, the actual set of literals satisfies all the structural equations. By contrast, the set of literals $\{f, \neg l, r, \neg e\}$ does not satisfy $\mathrm{e} = \mathrm{l} \vee \mathrm{r}$. When plugging in the literals, the truth values of e and $\mathrm{l} \vee \mathrm{r}$ do not match. Finally, we say that a set V of literals satisfies a set M of structural equations iff V satisfies each member of M.

The structural equations and the literals determine which events occur and which do not occur in a causal model. This determination can be expressed by a relation of satisfaction between a causal model and a propositional formula.

Definition 2. $\langle M, V\rangle$ satisfies ϕ
$\langle M, V\rangle$ satisfies ϕ iff ϕ is true in all complete sets V^c of literals that extend V and satisfy M. A set V^c of literals is complete iff it completely specifies which events in the causal scenario occur.

If V is complete, this definition boils down to: $\langle M, V\rangle$ satisfies ϕ iff V satisfies ϕ, or V does not satisfy M. Provided V is complete, $\langle M, V\rangle$ satisfies at least one of ϕ and $\neg\phi$ for any formula ϕ.

Our analysis relies on causal models that contain no information as to whether or not the presumed cause and the putative effect occur. We say that a causal model $\langle M, V\rangle$ is *uninformative* about a formula ϕ iff $\langle M, V\rangle$ satisfies none of ϕ and $\neg\phi$. Note that $\langle M, V\rangle$ cannot be uninformative on any formula if V is complete.

Consider a causal model $\langle M, V\rangle$, where $M = \{\mathrm{e} = \mathrm{l} \vee \mathrm{r}\}$. $\langle M, V\rangle$ is uninformative on e for $V = \emptyset$. There are four complete extensions that satisfy M. One of these is $\{\neg l, \neg r, \neg e\}$. Hence, $\langle M, V\rangle$ does not satisfy e. Similarly, $\langle M, V\rangle$ does not satisfy $\neg e$. There is a complete extension of V that satisfies M but fails to satisfy $\neg e$: the set $\{l, r, e\}$ of literals, for example, but also the sets $\{l, \neg r, e\}$ and $\{\neg l, r, e\}$. The structural equation constrains the represented scenario to four possible cases. These cases are expressed by the complete sets of literals that satisfy M.

Why is $\langle M, V\rangle$ not uninformative on e for $V = \{r\}$? Well, there is no complete extension of V that satisfies the structural equation in M but fails to satisfy e. There are only two such complete extensions: $\{l, r, e\}$ and $\{\neg l, r, e\}$. If r remains in the set V of literals, e is determined independent of whether or not l occurs.

It remains to introduce interventions. Recall that a structural equation $\mathrm{p} = \phi$ determines the truth value of the variable p if certain variables q occurring in ϕ are given truth values by the literals in V. To represent an intervention that sets p to one of the truth values, we replace the equation $\mathrm{p} = \phi$ by the corresponding literal p or $\neg p$. We implement such interventions by the notion of a submodel. M_I is a submodel of M relative to a consistent set I of literals just in case M_I contains the literals in I and the structural equations of M for the variables which do not occur in I. In symbols,

$$M_I = \{(\mathrm{p} = \phi) \in M \mid p \notin I \text{ and } \neg p \notin I\} \cup I.$$

We denote interventions by an operator $[\cdot]$ that takes a model M and a consistent set I of literals, and returns a submodel. In symbols, $M[I] = M_I$. In the causal model $\langle M, V\rangle$, where $M = \{\mathrm{e} = \mathrm{l} \vee \mathrm{r}\}$, we may intervene on the structural equation by $\{\neg r\}$. This yields: $M[\{\neg r\}] = \{\neg r, \mathrm{e} = \mathrm{l} \vee \mathrm{r}\}$.

The above definition of satisfaction applies to causal models and causal submodels. To be explicit, the definition does not only capture the relation of a causal model $\langle M, V\rangle$ satisfying a formula ϕ, but also the relation of a causal submodel $\langle M_I, V\rangle$ satisfying such a formula.

We are now in a position to spell out our analysis in a more precise way. The key idea is as follows: for c to be a cause of e, there must be a causal model $\langle M, V'\rangle$ that is uninformative about c and e, while intervening by $\neg c$ determines $\neg e$ to be true. In more formal terms:

Definition 3. Cause (Preliminary)
Let $\langle M, V\rangle$ be a causal model such that V satisfies M. c is a cause of e relative to $\langle M, V\rangle$ iff

(C1) $\langle M, V\rangle$ satisfies c and e, and
(C2) there is $V' \subset V$ such that $\langle M, V'\rangle$ is uninformative on c $\vee$ e, while $\langle M[\{\neg c\}], V'\rangle$ satisfies $\neg e$.

$\langle M, V'\rangle$ being uninformative on c $\vee$ e entails that $\langle M, V'\rangle$ satisfies none of c, $\neg c$, e, and $\neg e$. In what follows, we leave the disjunction implicit and simply say that $\langle M, V'\rangle$ is uninformative on c and e.

II. SCENARIOS

In this section, we test our analysis of causation against causal scenarios. We lay out the structure of causal scenarios by neuron diagrams. "Neuron diagrams earn their keep," as Paul and Hall write, "by representing a complex situation clearly and forcefully, allowing the reader to take in at a glance its central causal characteristics."[6] We introduce simple neuron diagrams for which there is always a corresponding causal model. Our causal models, however, are not limited to the causal scenarios which can be expressed in our simple neuron diagrams.

A neuron diagram is a graph-like representation that comes with different types of arrows and different types of nodes. Any node stands for a neuron, which fires or else does not. The firing of a neuron is visualized by a gray-shaded node, the non-firing by a white node. For the scenarios to be considered, we need two types of arrows. Each arrow with a head represents a stimulatory connection between two neurons, each arrow ending with a black dot an inhibitory connection. Furthermore, we distinguish between normal neurons that need just one stimulation for becoming excited and stubborn neurons that require two stimulations. Normal neurons are visualized by circles, stubborn neurons by thicker circles. A neuron diagram obeys four rules. First, the temporal order of events is left to right. Second, a normal neuron will fire if it is stimulated by at least one and inhibited by none. Third, a stubborn neuron will fire if it is stimulated by at least two and inhibited by none. Fourth, a neuron will not fire if it is inhibited by at least one.

Typically, neuron diagrams are used to represent events and absences. The firing of a neuron indicates the occurrence of some event,

[6] L. A. Paul and Ned Hall, *Causation: A User's Guide* (Oxford: Oxford University Press, 2013), p. 10. This being quoted, there are some shortcomings of neuron diagrams. For details, consult Christopher Hitchcock, "What's Wrong with Neuron Diagrams?," in J. K. Campbell, M. O'Rourke, and H. Silverstein, eds., *Causation and Explanation* (Cambridge, MA: MIT Press, 2007), pp. 4–69.

and the non-firing indicates its non-occurrence. Recall that we analyze causation between token events relative to a causal model $\langle M, V\rangle$, where the causal model represents the causal scenario under consideration. We thus need a correspondence between neuron diagrams and causal models.

Here is a recipe to translate an arbitrary neuron diagram, as detailed here, into a causal model. Given a neuron diagram, the corresponding causal model can be constructed in a stepwise fashion: for each neuron n of the neuron diagram,

(i) assign n a propositional variable p.
(ii) If n fires, add the positive literal p to the set V of literals.
(iii) If n does not fire, add the negative literal $\neg p$ to V.
(iv) If n has an incoming arrow, write on the right-hand side of p's structural equation a propositional formula ϕ such that ϕ is true iff n fires.[7]

This recipe adds a positive literal p to the set V of literals for each neuron that fires, and a negative literal $\neg p$ for each neuron that does not fire. Then the neuron rules are translated into structural equations. One can thus read off a neuron diagram its corresponding causal model: if a neuron is shaded gray, p is in the set V of literals of the corresponding causal model; if a neuron is white, $\neg p$ is in V.

We add the following feature to neuron diagrams. Dotted nodes represent neurons about which there is no information as to whether or not they fire. In more formal terms, if $p \notin V$ and $\neg p \notin V$, the corresponding neuron will be dotted. We portray now how our analysis solves the problems posed by switches, conjunctive causes, early and late preemption, prevention, and two scenarios of double prevention.

II.1. Switches. Recall the simple switch scenario whose neuron diagram is depicted in Figure 1. Here is a story that matches the neuron diagram: Flipper is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch (f) so that the train travels down the right-hand track (r), instead of the left (l). Since the

[7] The structural equations can be explicitly constructed from the rules governing neuron diagrams. That is, the catch-all condition (iv) can be replaced by the following clauses. (v) For each stimulatory arrow ending in a normal neuron n, add disjunctively to the right side of p's structural equation the variable that corresponds to the neuron where the arrow originates. (vi) For each pair of stimulatory arrows ending in a stubborn neuron n, add disjunctively to the right side of p's structural equation the conjunction of the two variables that correspond to the two neurons where the arrows originate. (vii) For each inhibitory arrow ending in n, add conjunctively to the right side of p's structural equation the negation of the variable that corresponds to the neuron where the arrow originates. This translation shows that there is a principled transition from simple neuron diagrams to our causal models.

tracks reconverge up ahead, the train arrives at its destination all the same (e).[8] Flipping the switch is not a (difference-making) cause of the train's arrival. By assumption, 'the train arrives at its destination all the same' independent of the flipping. And, by contrast to scenarios of preemption, the lack of a net effect is not due to a backup process independent of the flipping. Hence, flipping the switch makes no difference to the train's arrival.

Our recipe translates the neuron diagram of the switch scenario into the following causal model $\langle M, V\rangle$:

$\mathrm{l} = \neg \mathrm{f}$
$\mathrm{r} = \mathrm{f}$
$\mathrm{e} = \mathrm{l} \vee \mathrm{r}$
$f, \neg l, r, e$

Relative to $\langle M, V\rangle$, f is not a cause of e. The reason is that there exists no causal model $\langle M, V'\rangle$ uninformative on e. Any complete extension of the empty set V' of literals that satisfies the structural equations of M contains e. In fact, there are only two complete extensions that satisfy the structural equations, namely the actual $\{f, \neg l, r, e\}$ and the non-actual $\{\neg f, l, \neg r, e\}$. The structural equations in M determine e no matter what. f makes no difference as to e.

II.2. Conjunctive Causes. In a scenario of conjunctive causes, an effect occurs only if two causes obtain. The following neuron diagram depicts a scenario of conjunctive causes:

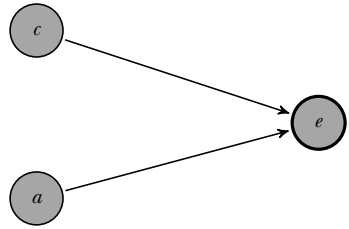


Figure 2.

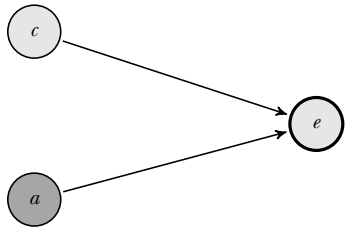
The neurons c and a fire. Together they bring the stubborn neuron e to fire. Had one of c and a not fired, e would not have been excited. Hence, the firing of both neurons is necessary for e's excitation.

[8] The example is taken from Hall's "Structural Equations and Causation," *op. cit.*, p. 28.

Our recipe translates the neuron diagram of Figure 2 into the following causal model $\langle M, V\rangle$:

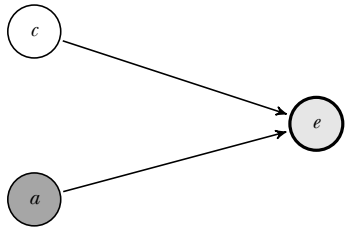
$e = c \wedge a$
c, a, e

Relative to $\langle M, V\rangle$, c is a cause of e. For this to be seen, consider the following causal model $\langle M, V'\rangle$ that is uninformative on c and e.



$e = c \wedge a$
a

Intervening by $\{\neg c\}$ yields:



$\neg c$
$e = c \wedge a$
a

This causal model determines $\neg e$ to be true. In more formal terms, $\langle M_{\{\neg c\}}, V'\rangle$ satisfies $\neg e$. Due to the symmetry of the scenario, a is a cause of e.

II.3. Early Preemption. Preemption scenarios are about backup processes: there is an event c that, intuitively, causes e. But even if c had not occurred, there is a backup event a that would have brought about e. The following neuron diagram represents the canonical structure of early preemption:

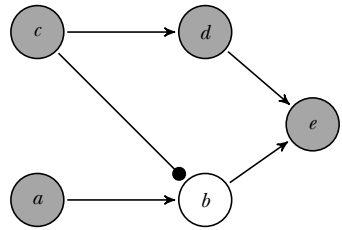


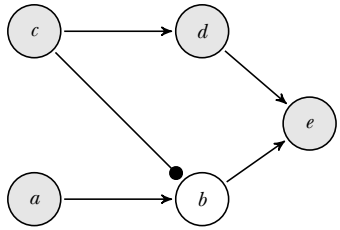
Figure 3.

c's firing excites neuron *d*, which in turn leads to an excitation of neuron *e*. At the same time, *c*'s firing inhibits the excitation of *b*. Had *c* not fired, however, *a* would have excited *b*, which in turn would have led to an excitation of *e*. The actual cause *c* preempts the mere potential cause *a*.

Our recipe translates the neuron diagram of early preemption into the following causal model $\langle M, V \rangle$:

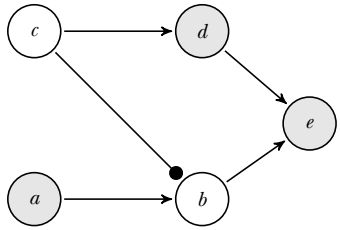
$\mathrm{d} = \mathrm{c}$
$\mathrm{b} = \mathrm{a} \wedge \neg \mathrm{c}$
$\mathrm{e} = \mathrm{d} \vee \mathrm{b}$
$c, a, d, \neg b, e$

Relative to $\langle M, V \rangle$, *c* is a cause of *e*. For this to be seen, consider the following causal model $\langle M, V' \rangle$ that is uninformative on c and e.



$\mathrm{d} = \mathrm{c}$
$\mathrm{b} = \mathrm{a} \wedge \neg \mathrm{c}$
$\mathrm{e} = \mathrm{d} \vee \mathrm{b}$
$\neg b$

Intervening by $\{\neg c\}$ yields:



$\neg c$
$\mathrm{d} = \mathrm{c}$
$\mathrm{b} = \mathrm{a} \wedge \neg \mathrm{c}$
$\mathrm{e} = \mathrm{d} \vee \mathrm{b}$
$\neg b$

This causal model determines $\neg e$ to be true. In more formal terms, $\langle M_{\{\neg c\}}, V' \rangle$ satisfies $\neg e$.

Relative to $\langle M, V \rangle$, a is not a cause of e. There is no causal model $\langle M, V' \rangle$—where $V' \subset V$—uninformative on a and e in which intervening by $\neg a$ would determine $\neg e$. Since the causal model is required to be uninformative on e, c cannot be in V'. But then there is the complete set $\{c, \neg a, d, \neg b, e\}$ of literals extending V' and satisfying the structural equations. In more formal terms, there is no $V' \subset V$ such that $\langle M, V' \rangle$ is uninformative on a and e, while $\langle M_{\{\neg a\}}, V' \rangle$ satisfies $\neg e$.

II.4. Late Preemption. Lewis subdivides preemption into early and late.[9] Figure 3 in the previous section depicts the canonical scenario of early preemption. There, the process started by the backup cause a is cut off before the process started by the preempting cause c has gone to completion. This means the process from the mere potential cause a is cut short at b before the effect e occurs. In scenarios of late preemption, by contrast, the backup process is cut off by the process of the actual cause running to completion. a is preempted only because c brought about the effect before a could do so.

Here is a story for late preemption.[10] Billy and Suzy are throwing rocks at a bottle. Suzy throws an instant earlier so that her rock hits the bottle first. Hence, Suzy's throw is the genuine cause of the bottle's shattering. Billy, however, is also very skillful at throwing rocks. If Suzy had not thrown her rock, Billy's rock would have hit the bottle, and thus the bottle would have shattered an instant later. Billy's throw is a preempted cause of the shattering of the bottle. The backup process initiated by Billy's throw is cut short only by Suzy's throw shattering the bottle. Crucially, the backup process is preempted only because the process starting from Suzy's throw runs to completion, and so brings about the shattering of the bottle before the backup process could do so. Until the bottle shatters there is always a backup process that would bring about this effect an instant later.[11]

[9] See David Lewis, "Postcripts to 'Causation'," in *Philosophical Papers, Volume II* (New York: Oxford University Press, 1986), pp. 172–213, here p. 200.

[10] Lewis provides a similar story for late preemption in his "Causation as Influence," *op. cit.*, p. 184.

[11] The problem posed by late preemption can be solved by fine-grained individuation conditions for events. According to these conditions, the shattering of the bottle and the shattering of the bottle an instant later are two different events. By adopting this strategy counterfactual accounts run into the trouble of spurious causation: they identify causal relations where, intuitively, there are none. See, for instance, Lewis's "Postscripts to 'Causation'," *op. cit.*, pp. 204–05, and ch. 3.4.2 of Paul and Hall's *Causation, op. cit.*

Lewis represents late preemption by a neuron diagram similar to the following.[12]

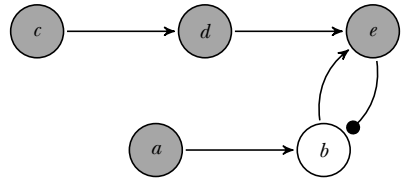


Figure 4.

Suzy throws her rock (c) an instant earlier than Billy does (a). Suzy's rock hits the bottle (d), and so the bottle shatters (e). The shattering of the bottle prevents Billy's rock from hitting the bottle ($\neg b$). The occurrence of the effect e cuts off the backup process started by a.

Earlier we have said that the temporal order of events in a neuron diagram is left to right. In Figure 4, however, the bottle shatters *first* (e) and so prevents that Billy's rock hits it ($\neg b$). In the non-actual scenario, where Suzy does not throw ($\neg c$), Billy's rock hits the bottle (b) *before* the bottle shatters (e). As is usual, the neuron diagram contains information about the actual and non-actual scenarios. As is unusual, merging the actual and non-actual scenarios in one neuron diagram violates here the rule of temporal order. In the actual scenario, $\neg b$ becomes actual after e occurs; in the non-actual scenario, b occurs before e does.[13]

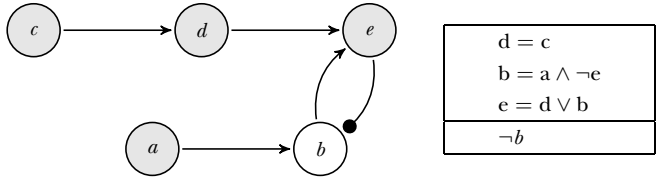
Our recipe translates the neuron diagram of late preemption into the following causal model $\langle M, V \rangle$:

$\mathrm{d} = \mathrm{c}$
$\mathrm{b} = \mathrm{a} \wedge \neg \mathrm{e}$
$\mathrm{e} = \mathrm{d} \vee \mathrm{b}$
$c, a, d, \neg b, e$

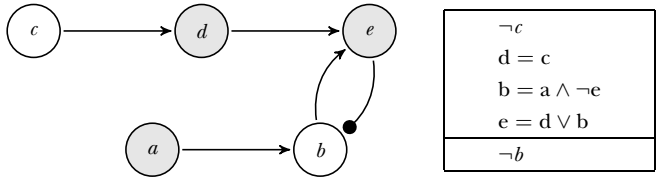
[12] See Lewis, "Postscripts to 'Causation'," *op. cit.*, p. 204.

[13] It is contested whether Lewis's story of Suzy and Billy is canonical for scenarios of late preemption. And even if it is, there is no unanimity on how to represent late preemption in neuron diagrams and causal models. Is the deviation from the rule of temporal order justified? Is the deviation justified when events are individuated in a relatively coarse-grained way as to their occurrence in time? Is Billy's rock *not* hitting the bottle (because there is no bottle any more) the absence to Billy's rock hitting the bottle (because Suzy's does not)? For discussions and a variety of tentative answers, see L. A. Paul, "Problems with Late Preemption," *Analysis*, LVIII (1998): 48–53; Hall, "Structural Equations and Causation," *op. cit.*; Christopher Hitchcock, "Prevention, Preemption, and the Principle of Sufficient Reason," *The Philosophical Review*, CXVI (2007): 495–532; and Paul and Hall's ch. 3.4 of *Causation*, *op. cit.*

Relative to $\langle M, V\rangle$, c is a cause of e. For this to be seen, consider the following causal model $\langle M, V'\rangle$ that is uninformative on c and e.



Intervening by $\{\neg c\}$ yields:



This causal model determines $\neg e$ to be true. In more formal terms, $\langle M_{\{\neg c\}}, V'\rangle$ satisfies $\neg e$.

Relative to $\langle M, V\rangle$, a is not a cause of e. The causal model $\langle M, V'\rangle$ is uninformative on a and e only for $V' = \emptyset$ or $V' = \{\neg b\}$. Intervening by $\{\neg a\}$ in each of them does not determine $\neg e$. For there is the complete set $\{c, \neg a, d, \neg b, e\}$ of literals which extends V' and satisfies the structural equations. Just like in the case of early preemption, there is no causal model $\langle M, V'\rangle$—where $V' \subset V$—uninformative on a and e in which intervening by $\neg a$ would determine $\neg e$. In more formal terms, $\langle M_{\{\neg a\}}, V'\rangle$ does not satisfy $\neg e$.

It should be mentioned that Halpern and Pearl represent the story of Suzy and Billy by an acyclic causal model.[14] Their causal model of late preemption can be obtained from the above causal model by replacing the equation for b with $b = a \wedge \neg d$. Our analysis delivers the desired results for this causal model as well. To sum up, our analysis solves late preemption—with respect to both Lewis's and Halpern and Pearl's representation of the scenario. Moreover, our analysis solves early and late preemption in a uniform manner.

[14] See Halpern and Pearl, "Causes and Explanations," *op. cit.*, pp. 861–63. Their causal model may well be preferable over the one derived from Lewis's neuron diagram. In the latter it is not well defined whether e is firing if a is and c is not. Relatedly, for $V_1 = \{\neg c, a\}$ there is no complete set V_1^c of literals that satisfies M and extends V_1. In the rock-throwing example of late preemption, by contrast, if Billy throws his rock and Suzy does not, the bottle *does* shatter.

II.5. Prevention. To prepare ourselves for a discussion of double prevention, let us take a look at simple prevention first. The basic scenario of prevention can be represented by the following neuron diagram:

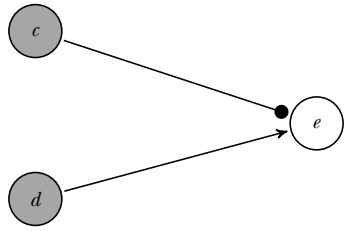


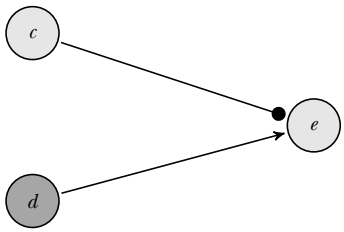
Figure 5.

Neuron c fires and thereby inhibits that neuron e gets excited. e would have been excited by d if the inhibitory signal from c were absent. But as it is, c prevents e from firing. That is, c causes $\neg e$ by prevention.

Our recipe translates the neuron diagram of prevention into the following causal model $\langle M, V\rangle$:

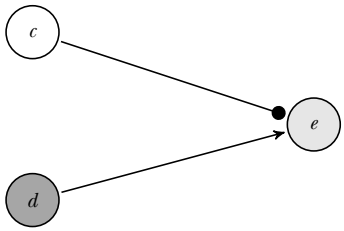
$\mathrm{e} = \neg\mathrm{c} \wedge \mathrm{d}$
$c, d, \neg e$

Relative to $\langle M, V\rangle$, c is a cause of $\neg e$. For this to be seen, consider the following causal model $\langle M, V'\rangle$ that is uninformative on c and e.



$\mathrm{e} = \neg\mathrm{c} \wedge \mathrm{d}$
d

Intervening by $\{\neg c\}$ yields:



$\neg c$
$\mathrm{e} = \neg\mathrm{c} \wedge \mathrm{d}$
d

This causal model determines e to be true. In more formal terms, $\langle M_{\{\neg c\}}, V' \rangle$ satisfies e. Moreover, d is not a cause of $\neg e$ relative to $\langle M, V \rangle$. Any causal model $\langle M, V' \rangle$ uninformative on d and e must be uninformative on c as well. Intervening by $\neg d$ in $\langle M, V' \rangle$ determines $\neg e$, and so does not determine e.

II.6. Double Prevention. Double prevention can be characterized as follows. c is said to double prevent e if c prevents an event that, had it occurred, would have prevented e. In other words, c double prevents e if c cancels a threat for e's occurrence. The characteristic structure of double prevention can be represented by the following neuron diagram:

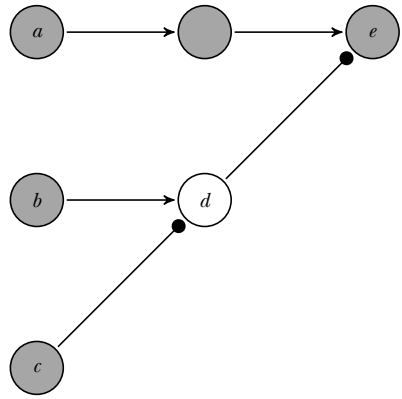


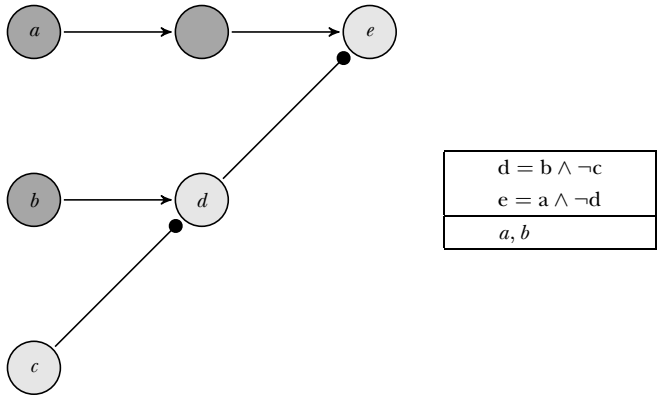
Figure 6.

c's firing prevents d's firing, which would have prevented e's firing. The example of double prevention exhibits a counterfactual dependence: given that b fires, e's firing counterfactually depends on c's firing. If c did not fire, d would fire, and thereby prevent e from firing. Hence, c's firing double prevents e's firing in Figure 6. In other terms, c's firing cancels a threat for e's firing, namely the threat originating from b's firing.

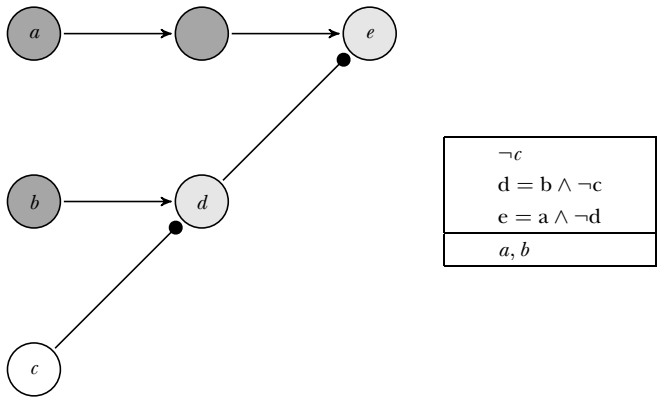
Our recipe translates the neuron diagram of double prevention into the following causal model $\langle M, V \rangle$:

$\mathrm{d} = \mathrm{b} \wedge \neg \mathrm{c}$
$\mathrm{e} = \mathrm{a} \wedge \neg \mathrm{d}$
$a, b, c, \neg d, e$

Relative to $\langle M, V\rangle$, c is a cause of e. For this to be seen, consider the following causal model $\langle M, V'\rangle$ that is uninformative on c and e.



Intervening by $\{\neg c\}$ yields:



This causal model determines d and so $\neg e$ to be true. In more formal terms, $\langle M_{\{\neg c\}}, V'\rangle$ satisfies $\neg e$.

II.7. Extended Double Prevention. A challenge to many counterfactual accounts of causation is an extension of the double prevention scenario depicted in Figure 6.[15] The extended version fits the structure of the following neuron diagram:

[15] This challenge has been presented in Hall's "Two Concepts of Causation," *op. cit.*, p. 247.

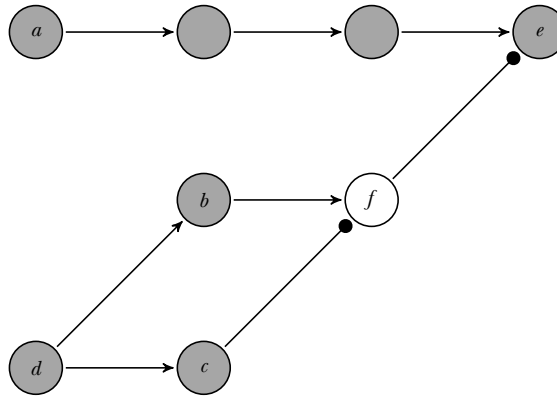


Figure 7.

Figure 7 extends Figure 6 by neuron d, which figures as a common cause of b and c. d starts a process via b that threatens to prevent e. At the same time, d initiates another process via c that prevents the threat. d cancels its *own* threat—the threat via b—to prevent e. In the example of the previous section, the threat originated independent of its preventer. Here, by contrast, d creates and cancels the threat to prevent e. This difference is sufficient for d not to be a cause of e.[16] Observe that the structure characteristic of double prevention is embedded in Figure 7. The firing of neuron c inhibits f's firing that, had it fired, would have inhibited e's firing. Nonetheless, this scenario of double prevention exhibits an important difference to its relative of the previous section: e does not counterfactually depend on d. If d had not fired, e would still have fired.

Here is a story that matches the structure of the scenario.[17] A hiker is on a beautiful hike (a). A boulder is dislodged (d) and rolls toward the hiker (b). The hiker sees the boulder coming and ducks (c) so that he does not get hit by the boulder ($\neg f$). If the hiker had not ducked, the boulder would have hit him, in which case the hiker would not have continued the hike. Since, however, he was clever enough to duck, the hiker continues the hike (e).

Hall calls the subgraph $d - b - c - f$ a *short circuit* with respect to e: the boulder threatens to prevent the continuation of the hike, but provokes an action that prevents this threat from being effective.[18]

[16] Or so argue Paul and Hall in *Causation*, *op. cit.*, p. 216.

[17] A similar story is presented in Hitchcock's "The Intransitivity of Causation Revealed in Equations and Graphs," *op. cit.*, p. 276.

[18] Hall, "Structural Equations and Causation," *op. cit.*, p. 36.

Like switching scenarios, the scenario seems to show that there are cases where causation is not transitive: the dislodged boulder (d) produces the ducking of the hiker (c), which in turn enables the hiker to continue the hike (e). But it is counterintuitive to say that the dislodging of the boulder (d) causes the continuation of the hike (e). After all, the dislodgement of the boulder has no net effect on the continuation of the hike and, by contrast to scenarios of preemption, there is no backup process in place—independent of the dislodged boulder—that would bring about the continuation of the hike anyway.

Our recipe translates the neuron diagram of the boulder scenario into the following causal model $\langle M, V\rangle$:

$\mathrm{b} = \mathrm{d}$
$\mathrm{c} = \mathrm{d}$
$\mathrm{f} = \mathrm{b} \wedge \neg \mathrm{c}$
$\mathrm{e} = \mathrm{a} \wedge \neg \mathrm{f}$
$a, d, b, c, \neg f, e$

Relative to $\langle M, V\rangle$, d is not a cause of e. The reason is that the causal model $\langle M, V'\rangle$ is only uninformative on d and e for $V' = \emptyset$. But $\langle M_{\{\neg d\}}, V'\rangle$ does not satisfy $\neg e$ for $V' = \emptyset$. In words, the causal model $\langle M, V'\rangle$ is uninformative about d and e only if the set of literals is empty. But then intervening by $\neg d$ does not make $\neg e$ true.

Note that a is necessary for determining e. If we were to keep a in the literals, the model would not be uninformative. There is no complete extension of $V' = \{a\}$ that satisfies all the structural equations of M but fails to satisfy e. Observe the similarity of the short circuit $c - b - d - f$ to the switching scenario in Figure 1. There, f does not make a difference as to e. Here, d does not make a difference as to f. Rather d defuses its own effects.

III. SYMMETRIC OVERDETERMINATION AND FINAL ANALYSIS

Our preliminary analysis cannot solve the scenario of symmetric overdetermination. Such scenarios are commonly represented by the following neuron diagram:

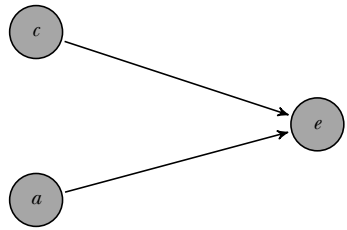


Figure 8.

Here is a story that fits the structure of overdetermination. A prisoner is shot by two soldiers at the same time (c and a), and each of the bullets is fatal without any temporal precedence. Arguably, both shots should qualify as causes of the death of the prisoner (e).

Our recipe translates the neuron diagram of Figure 8 into the following causal model $\langle M, V\rangle$:

$e = c \vee a$
c, a, e

The scenario of overdetermination differs from the scenario of conjunctive causes only in the structural equation for e. While the structural equation is conjunctive in the scenario of conjunctive causes, here the equation is disjunctive. The occurrence of one of the events, c or a, is sufficient for e to occur. But this means that each of c and a individually do not make a difference to e. Had c not fired, e would have fired nonetheless.

If c is not a difference maker as to e, should c count as a cause of e? If not, our preliminary analysis gives the right result. $\langle M, V'\rangle$ is uninformative on e only for $V' = \emptyset$. Intervening by $\neg c$ does not determine $\neg e$. After all, there is the complete set $\{\neg c, a, e\}$ of literals which satisfies the disjunctive structural equation.

But what caused the death of the prisoner? It seems as if we do not want to say that the death is uncaused. There is rather some agreement that each of the soldiers caused the death of the prisoner. We account for this intuition as follows: c is a cause of e because the factor comprising c and a makes a difference as to e. Each member of such a factor needs to be negated in order to make a difference as to the effect under consideration. In formal terms, $\{\neg c, \neg a\}$ determines $\neg e$. On this picture, a cause is a part of a difference-making factor. In what follows, C' stands for a difference-making factor and $\neg C'$ for the set containing all the negated elements of C'.

To state our final analysis of causation, we lift the restriction of cause and effect to single literals. A candidate for a cause is a set C

of literals, a candidate for an effect an arbitrary Boolean formula ε. Where C is a set of literals, $\bigwedge C$ stands for the conjunction of all literals in C, and $\bigvee C$ for the disjunction of all those literals.

Definition 4. Cause
Let $\langle M, V\rangle$ be a causal model such that V satisfies M. C is a cause of ε relative to $\langle M, V\rangle$ iff there is a superset C' of C such that

(C1) $\langle M, V\rangle$ satisfies $\bigwedge C \wedge \varepsilon$, and
(C2) (i) there is $V' \subset V$ such that $\langle M, V'\rangle$ is uninformative on $\bigvee C' \vee \varepsilon$ and $\langle M_{\{\neg C'\}}, V'\rangle$ satisfies $\neg\varepsilon$, while
(ii) for all $C'' \subset C'$, there is no $V'' \subset V$ such that $\langle M, V''\rangle$ is uninformative on $\bigvee C'' \vee \varepsilon$ and $\langle M_{\{\neg C''\}}, V''\rangle$ satisfies $\neg\varepsilon$.[19]

Let us explain condition (C2). The basic idea is still that a cause makes a difference to an effect in a causal model uninformative on the cause and the effect. A cause is now understood as a part of a difference-making factor C'. In fact, the first condition of (C2) simply demands that a genuine cause C is a subset of some difference-making factor C'. The second condition of (C2) ensures that the difference-making factor C' is minimal relative to the effect and the causal model uninformative on this effect: all strict subsets C'' of C' do *not* make a difference as to whether the effect is actual. This minimality condition excludes causally non-relevant events from being causes.

Our final analysis solves the scenario of symmetric overdetermination. Relative to $\langle M, V\rangle$, $C = \{c\}$ is a cause of e. For (i) there is a factor $C' = \{c, a\}$ which makes a difference with respect to the effect e relative to the causal model $\langle M, \emptyset\rangle$, which is uninformative on each proposition occurring in C' and e. And (ii) there is no strict subset of C' that makes a difference as to e relative to $\langle M, \emptyset\rangle$. Due to the symmetry of the scenario, a is a cause of e. (Moreover, the set $\{c, a\}$ counts as a cause of e.)

Symmetric overdetermination has troubled counterfactual accounts since Lewis's first analysis. Each of the overdetermining causes individually does not make any difference to the occurrence of the effect. Indeed, our analysis can capture overdetermination only by shifting the focus on the parts of a minimal difference-making factor. On the positive side, we are not aware of any other counterfactual analysis of causation which captures overdetermination, conjunctive causes, early and late preemption, switches, prevention, double prevention, and extended double prevention.

[19] If one wants cause and effect to be distinct, one should amend the definition by a clause like this: no element of C' occurs in ε.

IV. CONCLUSION

We have put forth an analysis of causation. In essence, c is a cause of e just in case c and e are actual, and there is a causal model uninformative on c and e in which c makes a difference as to e. Our final analysis successfully captures various causal scenarios, including overdetermination, preemption, switches, and extended double prevention. The competing counterfactual accounts of causation fail to capture—to the best of our knowledge—at least two of the causal scenarios considered. With respect to this set, our analysis is strictly more comprehensive than other counterfactual accounts.

The following table compares the results of our analysis (DMC) to the counterfactual accounts of Lewis ($\mathcal{L}$'73), Hitchcock (Hitch'01), Halpern and Pearl (HP'05), and Halpern (H'16).

Causes of e or $\neg e$	$\mathcal{L}$'73	Hitch'01	HP'05	H'16	DMC
Overdetermination	–	c, a	c, a	$\{c, a\}$	$c, a, \{c, a\}$
Conjunctive Causes	c, a	c, a	c, a	c, a	c, a
Early Preemption	c	c	c	c	c
Switch	f	f	f	f	–
Prevention	c	c	c	c	c
Double Prevention	c	c	c	c	c
E. Double Prevention	d	d	d	d	–

We think our results are what a notion of difference-making causation should deliver.

HOLGER ANDREAS

University of British Columbia

MARIO GÜNTHER

The Australian National University
LMU Munich