Paul Noordhof, *A Variety of Causes*, Oxford University Press, 2020, 592 pp.

Review by J. Dmitri Gallow | commissioned by *Notre Dame Philosophical Reviews*

Paul Noordhof's *A Variety of Causes* presents and defends a counterfactual theory of causation. The book is *incredibly* detailed; Noordhof dots every possible "i" and crosses every possible "t". It contains an extended discussion of Humean supervenience, dives deep into the theory of counterfactuals and the metaphysics of events which the theory presupposes, contains detailed discussion of so-called 'negative' causation, causal processes, the non-symmetry of causation, the relationship between causation and agency, causation and laws of nature, the metaphysics of chance, and much else—more than I can concisely list in this review. Readers interested in any of these topics will find stimulating discussion in *A Variety of Causes*—though they may find the discussion somewhat daunting to consume. If I were to raise one concern with the book, it would be that it does little to assist readers who wish to 'skip ahead' to a certain topic without taking in the entirety of the vast theoretical apparatus developed over 500 pages, and without reading through detailed exegesis and criticism of alternative proposals from the literature. While the discussion is rich, it is not bite-sized.

In this review, I will focus on the theory of causation which forms the heart of the book, and which is presented and defended in chapter 4. According to Noordhof, causation is not counterfactual dependence, but rather counterfactual dependence *modulo* a set of possible events, Σ. Some notation: let '*C*' and '*E*' be events which actually occurred, and let '*c*' and '*e*' be the propositions that *C* and *E* happen, respectively. For any set of events Σ, let σ be the proposition that some event in Σ happens—so that $\neg\sigma$ is the proposition that no event in Σ happens. Noordhof's full definition of Σ-dependence has bells and whistles to handle the nuances of probabilistic causation, but if we focus on deterministic systems, this definition can be simplified. In that special case, we can say that

> *Σ-Dependence (Determinism):* In deterministic systems, for any set of possible events, Σ, *E* Σ-depends upon *C* ($C \notin \Sigma$) iff
>
> (i) If *C* were to occur without any of the events in Σ occurring, then *E* would occur
>
> $$(\neg\sigma \wedge c) > e$$
>
> and

(ii) If neither C nor any of the events in Σ were to occur, then E would not occur

$$(\neg\sigma \wedge \neg c) > \neg e$$

Causation is then defined as follows.

> *Causation*: C is a cause of E iff there is an appropriate set of possible events, Σ, such that
>
> (1) E Σ-depends upon C
> (2) there is no superset of Σ, Σ^*, such that both
> a. E Σ^*-depends upon C, and
> b. there is a non-actual event (an event which did not actually happen), $N \notin \Sigma^*$, such that E Σ^*-depends upon N.

If the set of events Σ can be used to show that C is a cause of E in *Causation*, say that Σ is a *witness* to C causing E.

I say that Σ must be an "appropriate" witness (the term is mine, not Noordhof's). What it takes for Σ to be appropriate is specified on page 139—though I must confess that I am not sure I have understood the definition. Thankfully, I don't think we have to get distracted with the details. For Noordhof's goal in restricting the account to "appropriate" witnesses is to deal with situations in which the witnesses failing to occur would, *on its own*, change the times at which E may occur. So it appears that, so long as 'turning off' the events in Σ doesn't make any difference to when E might occur, Σ will be an appropriate witness.

We can sidestep issues having to do with indeterminism and issues related to the time at which the effect may happen by focusing on simple deterministic systems in which each event can only happen at a single time. So I'll limit my attention to system of neurons, connected by stimulatory and inhibitory synapses, where every neuron is only able to fire at a very specific time. For instance, consider the case of preemption shown in figure 1.

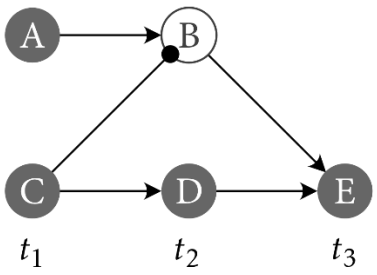

Figure 1: Preemption

In figure 1, each circle represents a neuron which can either fire or not fire at the time written underneath it. If a neuron fires at its designated time, then it is coloured grey. If it remains dormant, then it is coloured white. Arrows represent stimulatory synapses, whereas circular-headed connections (like the connection between C and B) represent inhibitory synapses. If C fires at t_1, then D will certainly fire at t_2, and B will certainly *not* fire at t_2—whether or not A fires. And E will fire at t_3 exactly if either B or D fires at t_2. (I am using non-italicised uppercase letters like 'N' to name neurons, and italicised uppercase letters like '*N*' for the event of the neuron N firing. Lowercase italicised letters like '*n*' therefore stand for the proposition that N fired.)

In *Preemption*, *C* is a cause of *E*. However, *A* is a pre-empted *backup* cause of *E*—were it not for *C*, *A* would have been a cause of *E*. The presence of this preempted backup backup means that *E* does not counterfactually depend upon *C*. For, were *C* to not happen, *A* would have caused *E* to happen. Nonetheless, Noordhof's counterfactual theory tells us that *C* is a cause of *E*, if we use the witness $\Sigma = \{ A \}$. For even though *E* does not depend upon *C*, it does $\{ A \}$-depend upon *C*,

$$(\neg a \wedge c) > e$$
$$(\neg a \wedge \neg c) > \neg e$$

And, moreover, there's no superset of Σ, Σ^*, such that *E* Σ^*-depends upon both *C* and some non-actual event $N \notin \Sigma^*$. So *Causation* rules that *C* is a cause of *E*.

To understand the reason for condition (2) in *Causation*, notice first that *A* *also* satisfies condition (1). For *E* $\{ C \}$-depends upon *A*,

$$(\neg c \wedge a) > e$$
$$(\neg c \wedge \neg a) > \neg e$$

Condition (2) uses the fact that there is a gap in the potential causal process leading from *A* to *E* to distinguish *A* from *C*. The gap comes when *B* does not occur. Because this gap exists, there is a non-actual event—namely *B*—upon which *E* $\{ C \}$-depends. That is,

$$(\neg c \wedge b) > e$$
$$(\neg c \wedge \neg b) > \neg e$$

So even though condition (1) is satisfied, condition (2) is not. So $\{ C \}$ is not a witness to *A* causing *E*. In general, for Σ to be a witness to *C* causing *E*, *E* must Σ-depend upon *C*, and must *not* Σ-depend upon any events which didn't actually occur.

But condition (2) of *Causation* doesn't just say that you can't have *E* Σ-depend upon a non-actual event. It says further that you can't have *E* Σ^*-depend upon a non-actual event, for any *superset* $\Sigma^* \supseteq \Sigma$. To appreciate why this additional strength is included, consider the system of neurons shown in figure 2.

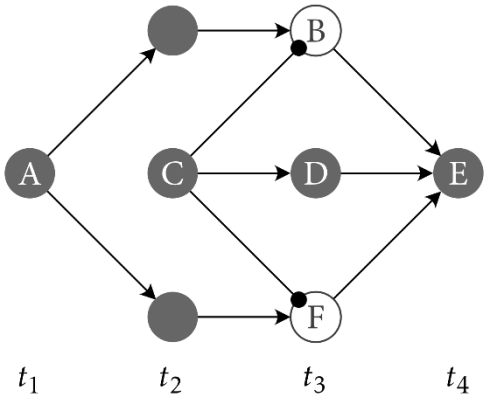

Figure 2: A is a backup which would have symmetrically overdetermined E, had C not fired.

In figure 2, there are two 'gaps' in the backup processes leading from *A* to *E*. That is: neither *B* nor *F* happen. And because *B* and *F* together symmetrically overdetermine *E*, were they to both happen, *E* does not { C }-depend upon either of these non-actual events individually. But Noordhof notes that *E does* { C, F }-depend upon the non-actual event *B*, and *E* does { C, B }-depend upon the non-actual event *F*. So the additional strength of condition (2) allows Noordhof to deal with cases of pre-emption like this as well.

It seems that the motivating idea behind Noordhof's theory is this: *C* can be a cause of *E* without *E* counterfactually depending on *C* when there are other, would-be causes. (A would-be cause is just something that would be a cause, were *C* to not happen.) Would-be causes can sever the tell-tale counterfactual relationship between *E* and *C*. But, if we 'turn off' these would-be causes of *E*, we should be able to restore counterfactual dependence between *E* and *C* without creating any new causal processes leading from *C* to *E*. Condition (1) of *Causation* is meant to require that *E* depends upon *C* when any would-be causes of *E* are 'turned off', and condition (2) is meant to require that 'turning off' these would-be causes of *E* doesn't create any *new* causal processes leading from *C* to *E*.

This approach to causation has much in common with the approach of authors like Hitchcock (2001), Woodward (2003), and Halpern & Pearl (2005). One important difference lies in the fact that, whereas these authors allow you to check for causation by checking for counterfactual dependence while 'holding fixed' whether certain other events occur or not, Noordhof only allows you to check for causation by checking for counterfactual dependence while 'holding fixed' the *non*-occurrence of events. In some cases, Noordhof's approach seems to do better. Consider, for instance, the neuron system shown in figure 3, which Hall (2007) calls a 'short circuit'.

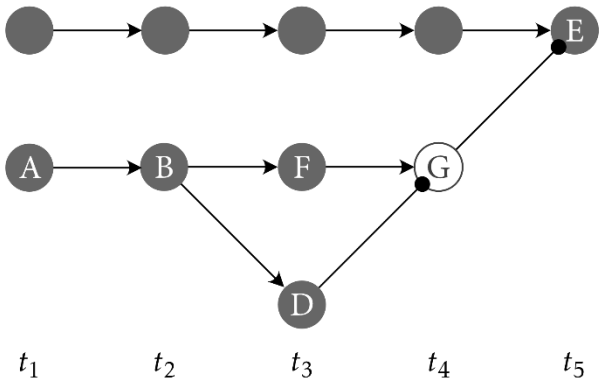

Figure 3: A 'short circuit'. Noordhof's theory says that neither A nor B is a cause of E.

If we 'hold fixed' the occurrence of *F*, then whether E fires will counterfactually depend upon whether B does. So the theories of Hitchcock, Woodward, and Halpern & Pearl will say that both *A* and *B* are causes of *E*. Noorhof's theory disagrees, denying that either *A* or *B* is a cause of *E*. This strikes many of us as the right verdict (though some disagree).

However, consider what Noordhof says about the neuron C in figure 4.

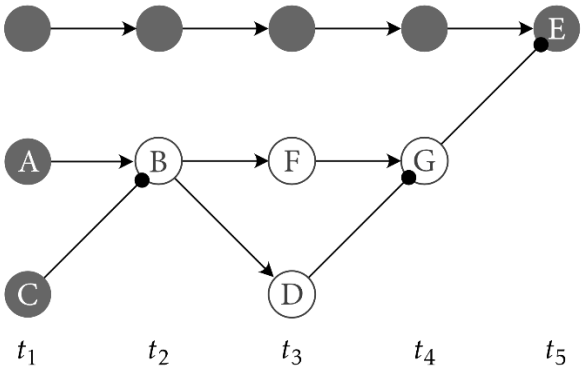

Figure 4: A prevented 'short-circuit'. Noordhof's theory says that C is a cause of E.

This neuron system is exactly like the one in figure 3, except that C prevents B from firing. In *this* case, *E* will counterfactually depend upon *C* when we hold fixed the *non*-occurrence of *D*. So Noordhof's theory will tell us that *C* is a cause of *E*.

More carefully, in figure 4, *E* { *D* }-depends upon *C*,

$$(\neg d \wedge c) > e$$
$$(\neg d \wedge \neg c) > \neg e$$

Moreover, the only relevant non-actual events not in $\Sigma = \{ D \}$ are *B, F,* and *G*. But, had any of those events occurred without *D*, they would have *prevented E* from occurring. So *E* does not { *D* }-depend upon any of those non-actual events. Nor does it Σ^*-depend upon *B*, *F*, or *G*, for any superset $\Sigma^* \supseteq \{D\}$.

Both intuitively and according to Noordhof's theory, *B* has no effect on *E* in figure 3. But Noordhof's theory tells us that, in figure 4, *C* causes *E* by preventing *B*. This seems like the wrong result. If *B* has no effect on *E* when it happens, and the only thing *C* does is prevent *B* from happening, then *C* should not count as a cause of *E*.

Or consider the neuron system shown in figure 5.

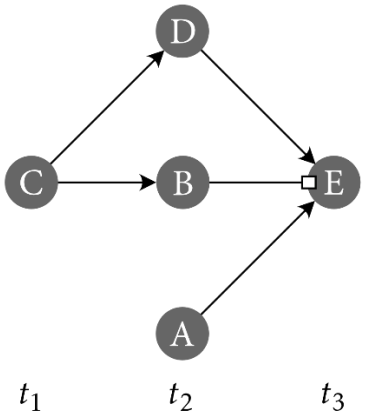

Figure 5: If B fires, E will need two stimulatory signals to fire. Noorhof's theory says that C is a cause of E.

Here, the connection between B and E is a *partially inhibitory* connection. If B fires, then E will need *two* stimulatory signals in order to fire. If, however, B doesn't fire, then E will only need one signal to fire. This neuron system seems similar to Hall's 'short circuit' from figure 3. In figure 3, *A* initiates a threat to *E* along one path by making F fire—which threatens to make G fire, which would prevent E from firing. But at the same time, *A* diffuses that very threat along another path by making D fire—thereby keeping G from preventing E from firing. Likewise, in figure 5, *C* initiates a threat to *E* along one path by making B fire—which threatens to keep E from firing. But at the same time, *C* diffuses that very threat along another path by making D fire—thereby keeping *B* from preventing E from firing. Just as I'm inclined to think that *A* is not a cause of *E* in figure 3, I'm inclined to think that *C* is not a cause of *E* in figure 5.

Despite their similarities, Noordhof's theory distinguishes the two cases. In figure 3, it says that *A* is not a cause of *E*. However, in figure 5, it says that *C is* a cause of *E*. For consider the witness $\Sigma = \{A, B\}$. E $\{A, B\}$-depends upon *C*,

$$(\neg a \wedge \neg b \wedge c) > e$$
$$(\neg a \wedge \neg b \wedge \neg c) > \neg e$$

and, since there are no non-actual events to be considered in this neuron system (every neuron fires), there's no superset $\Sigma^* \supseteq \{A, B\}$ such that E Σ^*-depends upon a non-actual event.

The reader is left wondering whether, according to this theory, counterfactual dependence between distinct events is sufficient for causation. To appreciate why this isn't clear, consider the neuron system shown in figure 6.

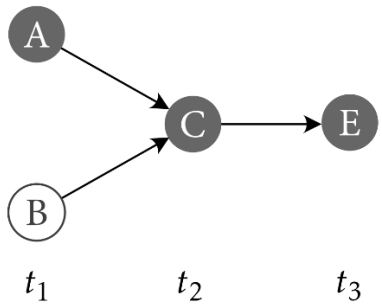

Figure 6: E ∅-depends upon C, but ∅ is not a witness to C causing E.

Here, *E* ∅-depends upon *C*. (That is to say: *E* counterfactually depends upon *C*. Were C to fire, so too would E; and were C to not fire, neither would E.) But the empty set is *not* a witness to *C* causing *E*. For there is a superset of ∅, namely { *A* }, such that (2a) *E* { *A* }-depends upon *C*, and (2b) there is a non-actual event not in { *A* }, namely *B*, such that *E* { *A* }-depends upon *B*. Were B to fire without A firing, E would fire; and, were neither A nor B to fire, E would not fire either.

$$(\neg a \wedge b) > e$$
$$(\neg a \wedge \neg b) > \neg e$$

Now, in this particular case, *Causation* does tell us that *C* is a cause of *E*. For, in this case, { *B* } is a witness to *C* causing *E*. But it is far from clear to this reader whether, in general, there will always be some set Σ which witnesses *C* causing *E* whenever *E* ∅-depends upon *C*. (Matters here are complicated by the fact that, in general, events can happen at different times, requiring us to attend to the complexities of Noordhof's criteria for what it takes for a witness to be "appropriate" on page 139.)

In sum: I have a few concerns and lingering questions about the theory of causation defended in chapter 4. But I should note that many of the book's theses and arguments are largely independent of the finicky details of chapter 4. The book's broad defence of a counterfactual approach to causation could be paired with any of a large number of counterfactual theories. Readers interested in counterfactual approaches to causation will find much to ponder over and learn from.

References

Hall, Ned. 2007. "Structural Equations and Causation." In *Philosophical Studies*, 132 (1): 109–136.

Halpern, Joseph Y. & Pearl, Judea. 2005. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." In *The British Journal for the Philosophy of Science*, 56: 843–887.

Hitchcock, Christopher. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." In *The Journal of Philosophy*, 98 (6): 273–299.

Woodward, James. 2003. Making Things Happen: A Theory of Causal Explanation. Oxford: Oxford University Press.