Causal Complexity, Conditional Independence, and Downward Causation

James Woodward*

This article defends the notion of downward causation, relating it to a notion of conditional independence.

1. Introduction. There are many possible strategies for dealing with causal complexity. I will focus on just one having to do with finding variables and theories for which a relationship I will call *conditional independence* (or *conditional irrelevance*) holds: this holds when a theory or explanation framed in terms of variables L of high dimensionality or many degrees of freedom that purports to account for some explanandum E can be replaced with a theory framed in terms of variables U with lower dimensionality or degrees of freedom, thus resulting (in this sense) in a reduction of complexity, as well as (often) gains in calculational or epistemic tractability.

I will illustrate this strategy via a discussion of downward causation (DC). DC has been criticized by many philosophers and some scientists as illegitimate or incoherent. I see it instead as a legitimate way of dealing with causal complexity. My plan is thus to begin with an account of DC, show how this applies to specific examples, and then broaden my discussion by connecting it to the notion of conditional independence.

2. Downward Causation. To make sense of the notion of DC we first need an account of causation. I adopt the interventionist account defended in Woodward (2003):

*To contact the author, please write to: History and Philosophy of Science, University of Pittsburgh, 101 Cathedral of Learning, 4200 Fifth Avenue, Pittsburgh, PA 15260; e-mail: jfw@pitt.edu.

Philosophy of Science, 87 (December 2020) pp. 857–867. 0031-8248/2020/8705-0007$10.00
Copyright 2020 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/710631 Published online by Cambridge University Press

> *X* causes *Y* if and only if there is some possible intervention that changes the value of *X* and under this intervention a regular change in the value of *Y* occurs.[1]

When *X* is at a "higher level" than *Y*, and this pattern of dependency obtains between interventions on *X* and *Y*, *X* downward causes *Y*.[2]

A very similar "interventionist" understanding of DC can be found in the recent scientific literature. Ellis (2016, 16) writes: "One demonstrates the existence of top-down causation whenever manipulating a higher-level variable can be shown to reliably change lower-level variables." To make this idea precise, some additional explication is required. Assume we are dealing with cases in which we have two sets of variables that I will label *U* (for upper) and *L* (for lower). Values of the *U* variables are "multiply realized" by values of the *L* variables, where realization is understood as a noncausal relation. As one way of making this precise, I adopt the following characterization: realization is a relationship between values of the *L* and *U* variables that can be represented by a many-one surjective function.[3] This includes (at least) the following two possibilities:

1. Two or more different lower-level variables are aggregated in some way (e.g., by addition) to form an upper-level variable: the value of a variable representing low density cholesterol (LDC) for some subject and the value for high density cholesterol (HDC) for that subject may be added to give a value for total cholesterol (TC) for the subject. Each pair of values of LDC and HDC with the same total "realizes" a single value of TC. Similarly given a set of lower-level variables, each of which represents the kinetic energy of different individual molecule in a gas, the average of these values realizes a value of the upper-level variable characterizing the average kinetic energy of the gas. Many

1. In order to avoid needless complexity, I will often describe causal relata as "variables," but readers should understand this as shorthand for "whatever in the world corresponds to variables."

2. The notion of "level" is used in many different and not entirely consistent ways in both science and philosophy. I think it doubtful that there is any single characterization that will fit all such uses. Rather than trying to provide one in the short space allotted, I rely instead on some generally accepted judgments about levels in the scientific literature. For example, I assume that variables are often legitimately regarded as at different levels when one is a coarse-graining of the other (e.g., via averaging or via some procedure such as addition that results in loss of degrees of freedom) and that variables used to characterize wholes are often legitimately regarded as at a level different from variables that characterize their parts.

3. I do not claim that all cases in which lower-level variables realize upper-level variables take this form. I adopt these assumptions because they enable precise discussion.

https://doi.org/10.1086/710631 Published online by Cambridge University Press

different combinations of values for the lower-level variable realize the same value of this upper-level variable.

2. Many values of a single lower-level variable may be collapsed into a single value for an upper-level variable: a variable measuring mass with a range of any positive real value is collapsed into a two-valued variable Z that takes values {mass $<$ 10 kg, mass $\geq$ 10 kg}. Any value of mass below 10 kg realizes the same value of the Z variable.

In both cases, the values of the Ls entirely fix the values of the Us. In both cases, a move from L to U will involve a reduction in dimensionality or degrees of freedom.

What we are interested in, then, is when if ever it is legitimate to treat an upper-level U, multiply realized by L, as a cause of a third variable Z that is distinct from U and L. In particular, how should we understand the notion of an intervention on U in this context? Whenever an upper-level U is realized by lower-level L, an intervention that changes the value of U to u will of course involve some change in some value of L, with different values of L, all of which realize $U = u$, occurring under different instances of this intervention.[4] We thus do not require that an intervention on U change the value of U while leaving values of L that realize that value of U unchanged (since we are assuming that this is impossible).[5]

We also impose the following requirement: when values of U are realized by a number of different values of L, an intervention on U with respect to some second variable Y that sets $U = u$ must have a uniform (or approximately uniform) effect on Y for all lower-level realizations of the value $U = u$.[6] This excludes so-called ambiguous manipulations in which the effect of $U = u$ on Y depends on how $U = u$ is realized. This condition is required for the notion of an intervention on U with respect to Y to be well defined. This requirement is closely related to the notion of conditional independence described below. Note that uniformity of effect is always characterized relative to an effect variable Y. It is true or very nearly true that all molecular realizations of a particular value of a temperature variable T of an ideal gas when set by interventions on T will have the same uniform effect on

4. In other words, the intervention does not determine which value of L realizes $U = u$.

5. More technically, in contexts in which realization is present, the requirement in Woodward (2003) that an intervention on U with respect to a second variable Y not affect Y via "off-path" variables—variables that affect Y via a causal path that does not go through U—should be understood in such a way that the variables L that realize U are not treated as "off-path" variables. In other words, the Ls are not treated as potential confounders for the $U \rightarrow Y$ relationship. Justification for this is provided in Woodward (2015). It is also provided by the consideration that one does not want to define the notion of intervention in such a way that unconfounded interventions on upper-level variables are impossible.

6. A similar requirement is imposed in Ellis (2016), 121.

https://doi.org/10.1086/710631 Published online by Cambridge University Press

pressure.[7] However, the effect of such interventions on some other microscopic variable Z may depend on the details of the way in which T is realized.

There are two other conditions on causation in general (not just DC) that play a role in my subsequent discussion: first, causal relata must be variables or values of variables (or more pedantically whatever in the world corresponds to variables and values). To anticipate my discussion below, quantities like voltage represented by variables can stand in causal relations, but things or entity-like structures such as neuronal membranes cannot stand in causal relationships since these do not correspond to variables. Second, causal relata must be "distinct" (sec. 3).

3. The Hodgkin-Huxley Model of the Action Potential as an Example of Downward Causation. There are many plausible examples of DC, explicitly described as such, in Noble (2006) and Ellis (2016), among others. For reasons of space I focus on just one example: the Hodgkin-Huxley (HH) model. This describes factors causally affecting the generic shape of the action potential generated by an individual neuron. The neuron is described as a circuit in parallel with the potential across the neuronal membrane functioning as a capacitor and with sodium and potassium ion channels with time and voltage dependent conductances g_{Na}, g_K and associated currents. Changes in the membrane potential over time cause changes in the channel conductances and other aspects of the behavior of the channels.

The ion channels are in an obvious sense "part" of the cell membrane (they are embedded in the membrane). This is at least part of what makes the example look like a plausible case of DC, with the upper-level membrane potential causally influencing the lower-level behavior of the ion channels and why the model is described in these terms by scientists like Noble (2006).

The HH model appears to describe a coherent and indeed empirically well-supported causal structure. Why then have critics thought that DC is objectionable? They have appealed to a number of considerations, but for reasons of space I focus on just one.[8] This objection is that DC involves a

7. "Very nearly true" here means true for all except a set of measure zero for such values.

8. Other objections to DC include considerations based on causal exclusion arguments and the claim that DC will often involve causal cycles, which are taken to be objectionable. Woodward (2015) responds to exclusion-based arguments. In connection with cycles, it is correct that systems in which DC is present will often involve cycles—e.g., in the case of the HH model, not only does V influence the channel conductances, but the conductances causally influence V. Although the issue deserves more attention than I can give it here, I claim that (i) cycles can sometimes be replaced by chains of time-indexed variables and that, even if they cannot, (ii) there is nothing incoherent or objectionable about many examples of cycles—cycles are extremely common in biological and social systems. Often they may be represented by directed graphs with an interventionist interpretation.

https://doi.org/10.1086/710631 Published online by Cambridge University Press

whole acting downward on its parts and that this is inconsistent with the requirement that cause and effect be suitably "distinct" from each other. (For arguments of this sort, see Craver and Bechtel [2007] and Heil [2017].)

I think this objection is misguided, basically because it fails to distinguish what the variables in the HH model describe which are *quantities* or *magnitudes* such as voltage or current rather than *things* or objects such as the neuronal membrane. The latter are not what these variables refer to. To put the point in an abstract way, suppose P is a spatial or temporal part of some whole W, and let X be some variable characterizing some feature of W and Y a variable characterizing some feature of P. Then (I claim) it is entirely possible for X and Y to be distinct in a way that allows X to cause Y despite the parthood relationship between P and W. In other words, even if P and W fail some appropriate test for distinctness, it does not follow that X and Y will fail such a test. For example, the potential difference V and the ionic conductances g_{Na}, g_K seem, intuitively, sufficiently distinct to stand in a causal relationship even though the ionic channels of which those conductances are predicated are part of the cell membrane. In particular V and g_{Na} do not seem to exhibit the kind of failure of distinctness that is present between, for example, saying "hello" and saying "hello" loudly (to take a standard philosophical illustration of failure of distinctness that precludes causation).

Making these intuitive claims more precise requires a criterion for the kind of distinctness among variables that makes them acceptable candidates for causal relata. The criterion I will employ is discussed in more detail in Woodward (2015), where it is called independent fixability (IF). According to IF, variables in a set **V** are suitably distinct if and only if it is "possible" to set each variable in **V** to each of its values via an intervention while also setting any other variable in **V** to each of its values via an intervention. "Possible" here includes logical or conceptual as well as causal possibility. Variables that are sufficiently distinct to stand in causal relationships must satisfy IF.

The variables in the HH model satisfy IF. First V is clearly manipulable in a way that is independent of the values taken by the ionic conductances. This does not require questionable judgments about possibility: it is shown by some of the experiments used by HH to establish their model. These involved use of a "voltage clamp" that enabled the experimenters to impose a stable potential difference (at various levels that they were able to choose) across the cell membrane in a way that depended only on the value set by the clamp. The clamp thus functioned as an intervention device, with the membrane potential difference fixed by the device rather than by such endogenous causes as the operation of the ion channels. This allowed the experimenters to investigate (and isolate) the effect of V on the ionic currents and the conductances in a way that confirmed the predictions of the HH model. Moreover, although HH lacked the technology to carry out such experiments, it is now possible to intervene by molecular means to alter the individual ionic channel currents and

https://doi.org/10.1086/710631 Published online by Cambridge University Press

conductances independently of V and to confirm that the effects of such interventions on V conform to the HH model.

The experiments just described show that the relation between V and the ionic conductances satisfy the requirements on DC described in section 2. Interventions on V change the ionic conductances in a uniform regular way that does not depend on the microlevel realizations of particular values of V. Moreover, V and the conductances are sufficiently distinct to stand in causal relationships.

4. An Objection. Even readers who are somewhat persuaded by the discussion in previous sections may nonetheless feel the pull of the following objection. In the case of the HH model (the objection goes), the neuron itself is composed of atoms and molecules that interact locally, mainly through the electromagnetic force. The membrane potential, the channel conductances, and so on, must be the upshot or resultant of complex patterns of interaction among these atomic and molecular constituents. Variable V, the channel conductances, and other variables in the HH model thus do not represent anything "over and above" these atomic constituents and their interactions. Why then should we attribute causal efficacy to the upper-level variables at all? Talk of causation by upper-level variables may be a useful way of talking that we are forced to use because of our computational and epistemic limitations, but this does not mean that causal relationships involving upper-level variables are truly "out there" in the world.

An adequate response to this objection needs to describe what worldly information about causal relations upper-level causal claims track and how this information connects to causal information provided by lower-level variables in a way that identifies the conditions under which use of the upper-level causal claims is legitimate. In what follows I attempt to do this via the notion of conditional independence, showing how true upper-level causal claims, including claims of DC, reflect facts about how the world is organized and not just convenient ways of thinking.[9]

5. Conditional Independence. Suppose we have a set of upper-level variables U that are related to a set of lower-level variables L via the realization relation described in section 2: the values of L realize values of U via many-one surjective functional relationships so that the Us are a coarse-graining of the Ls, with these being of much higher dimensionality than the Us. Suppose also that the Ls are (unconditionally) causally relevant (by the standard interventionist criterion of relevance) to E, which may be either upper or

9. The ideas about conditional independence that follow are similar to and have been heavily influenced by the ideas about learning causal macrovariables in Chalupka, Eberhardt, and Perona (2017). Related ideas can be found in Yablo (1992).

https://doi.org/10.1086/710631 Published online by Cambridge University Press

lower level. Assume also that the *U*s are causally relevant to *E*. Unconditional relevance of *L* (*U*) to *E* means there are some changes in the values of *L* (*U*) when produced by interventions that are associated with changes in *E*. Finally, let us say that a set of *L* variables is irrelevant to (independent of) *E conditional* on the *U* variables if the *L*s are unconditionally relevant to *E*, the *U*s are unconditionally relevant to *E*, and conditional on the values of the *U*s changes in the value of *L* produced by additional interventions consistent with these values for the *U*s are irrelevant to *E*.[10] My claim is that when this conditional independence relation holds we may legitimately treat the *U* variables as upper-level causes of *E*: the *U*s capture everything that makes a difference for *E* so that to the extent that causal explanation has to do with difference-making, the explanatory import of the *L*s for *E* can be entirely absorbed into the *U*s.

As an illustration, suppose the *U*s are thermodynamic variables like temperature and the *L*s realizing these specify the position and momentum of each of the individual molecules comprising the gas. Then conditional on some particular value *t* of the temperature variable *T* (for each such value) further variations in the value of the position and momentum variables for the individual molecules consistent with $T = t$ will make no difference or very nearly no difference for the values of thermodynamic variables like pressure. Similarly in the HH model, *V* will be a legitimate upper-level cause of the channel conductances to the extent that it is true as an empirical matter that further lower-level detail of a sort that might be captured by lower-level variables describing, for example, individual atoms and molecules is irrelevant to these explananda conditional on the values of *V*.

The condition just described, involving complete irrelevance of the lower-level variables to *E* conditional on the values of certain upper-level variables, is obviously a limiting case, although for reasons described below it is arguably not as rare as many philosophers suppose. Put in language more familiar to scientists, this condition corresponds to complete *separation of scales* or *decoupling* between the two sets of variables *U* and *L* with respect to *E*.

10. Conditional independence in this context is thus to be understood in terms of *interventionist counterfactuals* (Briggs 2012) rather than probabilistic conditional independence. The idea is that we fix the *U* variable or variables to some value(s) *u* via an intervention and then consider independent interventions that set the *L* variables to values that are consistent with this value of *u*. The test for conditional independence is then whether, for all values of *U* fixed by interventions, such independent interventions on *L* make no difference to *E*. In other words, we consider whether counterfactuals of the following form hold: if *L* were set to *l* via an intervention, where *l* is any value consistent with $U = u$, then if $U = u$, the value of *E* would be the same for all such values of *L* (i.e., *E* depends only on $U = u$). Conditional independence requires that this be true for all values of *U*.

https://doi.org/10.1086/710631 Published online by Cambridge University Press

The requirement of complete irrelevance may be relaxed in various ways. One possibility is that although there may be rare or exceptional values of the *L*s that are conditionally relevant to *E*, even given the values of the *U*s, this may not be true for most or "almost all" values of the *L*s—for most or almost all such values, the values of *L* are conditionally independent of *E*, given *U*. Another possibility is that conditional irrelevance holds for all values of *L* and *U* within a certain large interval, including those values most likely to occur (at least around here right now). Under such conditions it may be reasonable to employ the *U*s in lieu of the *L*s.

Several further remarks may help to elucidate how this condition works. First, notice that, as with the characterization of interventions, conditional irrelevance is characterized with respect to a particular explanandum *E* (or some specified set of these). It is entirely possible (even typical) for *L* to be irrelevant to *E* conditional on *U* but for this conditional irrelevance relation not to hold relative to some other explanandum E^*. If we wish to explain not the generic shape of the action potential but details of the shape and time course of particular potentials, information at a lower level than that described by the HH model will likely be required. Sometimes, however, we find that there is a substantial set or grouping of different explananda (although not all possible explananda) for all of which conditional independence holds for some set of upper-level variables. Thermodynamic variables like temperature, pressure, volume, and entropy have this feature. Such variables can be thought of as constituting a sort of autonomous domain or protectorate—we do not need to go outside of it to explain variables within this grouping.

A second point is that the conditional irrelevance of some *L* to *E* conditional on some *U* does not imply that *L* is unconditionally irrelevant to *E*. On the contrary, when *L* is conditionally irrelevant to *E*, it will be unconditionally relevant to *E*. Thus, the argument of this article does not imply "downward exclusion"—*U* being an upper-level cause of *E* is compatible with the *L* that realizes *U* also being such a cause.[11] I mention this because philosophers who are sympathetic to causation or explanation involving upper-level variables sometimes say things like "the lower-level details are irrelevant." Not so—the lower-level details influence whether the square peg fits into the hole, the shape of the action potential, and so on. The point these philosophers are trying to express is better made as a claim about conditional independence: the lower-level variables are relevant, but the upper-level variables absorb all of this relevant detail into some smaller set of variables or variables with smaller dimensionality.

11. For a response to the objection that this involves an objectionable kind of causal overdetermination, see Woodward (2015).

https://doi.org/10.1086/710631 Published online by Cambridge University Press

6. The Role of Epistemic and Calculational Limitations. On the story so far, conditional independence relations help to explain why it is legitimate to employ upper-level variables as causes, including those that figure in downward causal relationships. However, there is more to be said: because of our epistemic and calculational limitations, we often lack information about lower-level variables, a theory connecting these to explananda of interest, and the ability to calculate the consequences of the lower-level information for these explananda. In such cases, considerations based on conditional independence and considerations deriving from our epistemic and computational limitations can work together in a cooperative, mutually reinforcing way: if we are fortunate, then even if we cannot use the lower-level variables because of tractability considerations, conditional independence considerations can show that we can use upper-level variables instead without explanatory loss.

It is worth emphasizing that this is not an attempt to justify appeals to upper-level variables just by appeal to facts about human limitations, independently of what nature is like. That certain variables L are conditionally irrelevant to other variables E, given the values of other variables U is a fact about the world and not a fact about us or what we are interested in or able to do. (These are the worldly causal facts, referred to earlier, that upper-level causal claims track and that are "out there" or "objective.") Our calculational limitations come into the story only after the conditional independence facts are shown to license the use of upper-level variables.

We should note another distinctive feature of the conditional independence justification for the use of upper-level variables. It is common in the philosophical literature (e.g., Weslake 2010; Franklin-Hall 2016) for defenses of causal claims or explanations in terms of upper-level variables to attempt to show that such explanations are superior to explanations in terms of lower-level variables or at least that they are superior insofar as we abstract from "pragmatic" considerations having to do with human epistemic and calculational limitations.

The conditional independence justification does not commit us to this claim about the superiority of upper-level causal explanations. Again, what it attempts to do is to identify conditions under which it is permissible or legitimate to employ upper-level variables—permissible in the sense that this can be done without explanatory loss. This does not imply that an explanation in terms of lower-level variables (if we could produce one) would be inferior to an explanation in terms of upper-level variables. This last claim may well be correct, but if so, it will require arguments and considerations different from those given above. Conditional independence gives us a way of defending or justifying the use of upper-level causal claims, including those involving DC, without taking on the difficult burden of arguing that these are superior to in principle lower-level explanations.

https://doi.org/10.1086/710631 Published online by Cambridge University Press

7. How Common Is Conditional Independence? Finally, let me return to this issue, raised in passing above. I assume that, as an empirical matter, for most arbitrary sets of *U*s, *L*s, and *E*s conditional independence, or even approximate conditional independence, will fail. This does not in itself indicate anything about the usefulness of the notion—it merely reflects that "good" upper-level variables are hard to find—indeed, hard to find even given a lower-level ground truth from which the upper- level variables can be constructed.[12] The interesting question is the extent to which there are cases in which conditional independence or something like it does hold. I conclude with two such cases.

I have already mentioned that decoupling results in high energy physics. These show that the physics at certain energy/length scales is almost entirely independent of the physics at higher energy/shorter length scales—independent in the sense that almost any reasonable higher energy theory would generate the same lower-level behavior. This is naturally understood as a claim about conditional independence and is both a blessing and a curse. On the one hand, decoupling implies that phenomena at lower energy scales can be theorized about and understood independently of the details of what is going on at higher energy scales. This is a good thing if the theory at higher energy scales is unknown and experimentally inaccessible. On the other hand, this independence means that phenomena observed at the lower energy scale tell us little or nothing about what the correct theory might be at higher scales, other than that the higher energy theory must be consistent with the lower-level observations.

Another kind of case comes from biology. Organisms are often shaped by natural selection and by other selective processes such as learning to be sensitive only to ecologically significant properties of their environment where these are relatively coarse-grained from the point of view of lower-level theorizing. In such cases, variations in lower-level detail are likely to be conditionally irrelevant to various organism behaviors conditional on upper-level coarse-graining of this detail. It would make little sense for bodily responses of medium-sized organisms like ourselves to fearful and stressful stimuli to vary depending on the exact details of the molecular realization of those stimuli—it is the general fact that the stimulus is a large predator that is relevant. In such cases and for most of sensory processing we have screening off of lower-level detail by ecologically relevant upper-level variables with respect to behavioral responses.

Finally consider cases in which conditional on some *U*s, certain lower-level variables *L* remain relevant to some *E*. In such cases one hopes (and it may often be true) that one can find some other set of L^*s that in conjunction with the *U*s

12. That is, even given *L*s that are relevant to *E*, most ways of forming *U*s from the *L*s will not yield conditional independence with significant dimension reduction.

https://doi.org/10.1086/710631 Published online by Cambridge University Press

do satisfy conditional independence with respect to the original set of *L*s and that furthermore achieve significant dimension reduction and simplification. For example, "psychological" variables by themselves may not capture all that is relevant to some behavior (conditional independence fails with respect to some full set of neurobiological variables), but the psychological variables in conjunction with a very limited set of neurobiological variables may accomplish this—if we are fortunate we may not have to appeal to the full set of neurobiological variables and still less to information that goes all the way down to the level of individual molecules. In such cases, conditional independence can function as a goal that tells us what we are looking for.

REFERENCES

Briggs, Rachel. 2012. "Interventionist Counterfactuals." *Philosophical Studies* 160:139–66.
Chalupka, Krzysztof, Frederick Eberhardt, and Pietro Perona. 2017. "Causal Feature Learning: An Overview." *Behaviormetrika* 44:137–64.
Craver, Carl, and William Bechtel. 2007. "Top-Down Causation without Top-Down Causes." *Biology and Philosophy* 22:547–63.
Ellis, George. 2016. *How Can Physics Underlie the Mind? Top-Down Causation in the Human Context*. Berlin: Springer.
Franklin-Hall, Laura. 2016. "High-Level Explanation and the Interventionist's 'Variables Problem.'" *British Journal for the Philosophy of Science* 67:553–77.
Heil, John. 2017. "Downward Causation." In *Philosophical and Scientific Perspectives on Downward Causation*, ed. Michele Paoletti and Francesco Orilia, 42–53. New York: Routledge.
Noble, Denis. 2006. *The Music of Life*. Oxford: Oxford University Press.
Weslake, B. 2010. "Explanatory Depth." *Philosophy of Science* 77:273–94.
Woodward, James. 2003. *Making Things Happen*. New York: Oxford University Press.
———. 2015. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91:303–47.
Yablo, Stephen. 1992. "Mental Causation." *Philosophical Review* 101:245–80.

https://doi.org/10.1086/710631 Published online by Cambridge University Press