To appear in *Top-Down Causation and Emergence*, Eds. J. Voosholz and M. Gabriel (Springer)

Downward Causation Defended[1]

James Woodward
HPS, University of Pittsburgh

1. Introduction

It is an honor and a pleasure to contribute to this festschrift for George Ellis. I first became interested in the topic of downward causation as a result of conversations that I had many years ago with Roger Sperry when I was a postdoc at Caltech. I've always thought that there was something right in the basic idea but it has only been recently, partly as a consequence of reading work by Ellis (and others such as Denis Noble) as well as some philosophical criticisms of downward causation that struck me as misguided that I have thought that I might have something to say on this subject. The ideas that follow reflect the influence of Ellis and Noble as well as some recent developments in machine learning and computer science concerning forming macro-variables from more fine-grained realizing micro-variables (e.g. Chalupka et al., 2017)[2].

I begin, though, with some stage setting and methodological remarks. I'm a philosopher of science with an interest in methodology and in causal reasoning. I approach the issues around downward causation from that perspective, not that of metaphysician. Although I address some metaphysical arguments against the possibility of downward causation, my primary concerns are methodological: my goal is to try to understand what it is about certain systems that inclines a number of scientists to characterize their behavior in terms of downward causation, whether such characterizations are ever correct, and if so, in what circumstances. I thus proceed on the assumption that the metaphysical issues are not the only ones that deserve philosophical attention[3].

I also approach this subject from what I have elsewhere described as a functional perspective (Woodward, 2014, forthcoming): we should think about causal claims in terms of the goals and functions that we want to such claims to serve—in terms of what we want to *do* with such claims. The interventionist account of causation I describe below embodies this functionalist picture: the idea is that one important function of causal claims is to describe the results of manipulations or interventions. This leads to the way in which I frame the issues around downwards causation: these have to do roughly with whether interventions on upper-

[1] Thanks to Jan Voosholz for helpful comments on an earlier draft.

[2] For additional relevant work in machine learning and computer science on forming macro-variables from underlying micro-variables see Beckers and Halpern, 2019, Rubenstein et al. 2017.

[3] Contrary to the anonymous referee for this chapter, who claims that the metaphysical issues are the only ones that "count". For the role that this rhetorical strategy of dismissal of the non-metaphysical plays in contemporary philosophical discussion, see Woodward (2017)

level variables can systematically change lower-level variables and if so, under what conditions this is possible.

As I note in Section 3, there are many cases, both drawn from various sciences and from common-sense causal thinking that seem to be naturally described in terms of downward causation, understood as described above. I do *not* claim that such descriptions are correct merely because they seem natural or prima-facie plausible or fit with what various scientists have said about the examples. As I said above, I'm very aware that there are in-principle metaphysical objections (based on causal exclusion arguments, worries about violations of the causal closure of the physical and so on) to the very possibility of downward causation. At the same time, as a philosopher of science, I think that it is very much in order to explore what it is about these examples that has led many to think of them in terms of downward causation. (In other words, I assume that if there are any plausible cases of downward causation, these are the sorts of examples we should be looking at.) This functional orientation leads me to explore such questions as what the use might be of a notion of downward causation, why we might find it fruitful to operate with such a notion, what kind of evidence might persuade us that downward causation is present and so on. Of course if the notion is incoherent for metaphysical reasons, then the fact that we might *like* to think in terms of downward causation cannot show that that the notion is legitimate. But if the metaphysical objections can be disarmed and if we can provide a coherent account of what downward causation involves, why such a notion is a useful one and what sorts of situations are appropriately described in terms of this notion, this can provide a vindication of the notion. In any event this will be my strategy.

Metaphysicians sometimes accuse philosophers of science like me of conflating epistemology/methodology with metaphysics or illegitimately arguing from the former from the latter. They acknowledge that we have methods that may be interpreted as providing evidence for downward causation and that it may be "pragmatically useful" to think in terms of this notion, but insist that this shows nothing about whether downward causation is "real", ontologically or metaphysically speaking. This line of argument raises issues that I cannot fully address in this paper. I will say, however, that on the functional approach to causation (and to methodology more generally) that I favor, we should not expect methodology/epistemology and metaphysics/ontology (insofar as the latter has to do with what is "really out there") to come apart in this way. On a functional notion of causation—one that we can use— the causal relations that are out there, must be such that, at least in some range of cases, we can know whether they are present are not. Our account of the methodology/epistemology of causal reasoning should to this extent fit with the worldly structures associated with such relationships.

Consider, in this light, someone who holds that what causation "really is"[4], metaphysically speaking, has nothing to do with what is disclosed by controlled experiments (the experiments being "merely of epistemological significance") so that even if there if are experiments in which upper-level variables are manipulated with associated changes in lower variables this tells us nothing about whether downward causation is "real". I can envision two possible defenses for such a claim. The first is simply that the experiments in question don't really show the presence of downward causation, because when so interpreted they are defective in some way—e.g., they fail to control for confounders which should be controlled for (which is one way of interpreting causal exclusion arguments). I claim that the defender of downward causation has a good response to this sort of objection. (Section 7). The other possible response

[4] For more on this theme, see Woodward, forthcoming.

is that even if the experiments I interpret as showing downward causation are unimpeachable from the point of view of experimental design, showing that "downward causation is "real" present requires some more – that is, there is some thicker notion of causation ("real causation") that fails to be present in apparent downward causation cases, even in the presence of experimental results like those described above . Here I would challenge those inclined to this view to explain what this "something more" involves, how to detect when it is present, why it is a useful or appropriate to have a conception of causation that incorporates it, and how this conception excludes downward causation. One would also like an explanation of why experimentation fails to detect causal relations in cases involving relations between upper and lower level variables but (presumably) succeeds in other cases. It is not obvious how such an account might proceed.

2. Causation and Intervention in the Presence of Realizing Relations.

To develop an account of downward causation (or, more generally, causal claims in which the candidate causes are upper-level variables and the effects either lower or upper level) we first need to specify what we mean by "causation". I adopt the following version of an interventionist or manipulationist account defended in Woodward, 2003:

> (**M**) Where *X* and *Y* are variables, *X* causes *Y* iff there are some possible interventions that would change the value of *X* and if were such intervention to occur, a regular change in the value of *Y* would occur[5].

Woodward, 2003 provides a a technically precise characterization of "intervention" and similar notions are characterized in Pearl, 2000 and Spirtes, Glymour and Scheines, 2000. However it is important to understand that these characterizations were intended to apply to cases in which only causal relations among variables and not supervenience relations are present. Further clarification is required when we apply the notion of an intervention to cases in which supervenience relations are present—see below. As long as we are restricted to cases in which no supervenience relations are present, we may think of an intervention *I* on *X* with respect to *Y* as causing a change in the value of *X* that is of such a character that any change in *Y*, should it

[5] A couple of additional remarks: First, in order to avoid needless verbiage, I will usually describe causal relata as "variables" but of course readers should understand this as shorthand for "whatever in the world corresponds to variables or to variables taking certain values". Thus causal relata are features like mass and charge that may be possessed by systems in the world. Also, in order to simplify the discussion, I will confine myself to cases in which the causal relationships in which we are interested are deterministic. In my view, nothing fundamental changes when we consider indeterministic causal relations, except that "regular change" needs to be interpreted as something like "regular change in the probability distribution of *Y"*. Finally, the "regular change" requirement in **M**, which is imposed in Woodward, 2003, pp. 41-2, means simply that there must be some values of *X* such interventions that set those values are followed by regular or uniform responses in *Y.* This is fully compatible with there being other values of *X* for which this is not true. In other words the condition in **M** that there must be "some" (not necessarily all) values of *X* associated under interventions with changes in *Y* should be understood as requiring that *for those values* there should be a uniform response in *Y*.

occur, occurs only "through" *X*. Expressed slightly differently, an intervention *I* on *X* with respect to *Y* is an unconfounded change in *X*—unconfounded in the sense that *I* does not affect *Y* via any *causal* route[6] that does not go through *X*. Manipulations of putative effect variables in well-designed experiments, including those achieved in randomized controlled trials, are paradigm cases of interventions. The intention behind **M** is to capture the common sense idea that the mark of a causal relationship is that causes are potential "handles" for changing effects; causal relationships are those relationships that can be exploited "in principle" for manipulation and control, in the sense that if manipulating *X* would be in principle a way of manipulating *Y,* then *X* causes *Y*, and conversely.

My conception of downward causation simply applies this interventionist picture to the case in which *X* is at a "higher level"[7] than *Y*. In such a case when (and only when) *Y* changes in a regular manner under interventions on *X*, *X* is a downward cause of *Y*. It is worth emphasizing that this is a "thin" notion of causation, both metaphysically and otherwise. For *X* to cause *Y* it is not required that there be a continuous process running from *X* to *Y*, that *X* "transmit" energy or "biff" or "umph" to *Y* (or anything similar). Nor is it required that *X* and *Y* are variables that occur in some "fundamental" theory drawn from physics. Readers should thus keep in mind that when I talk about downward causation all that I mean by causation is a relationship that satisfies **M** (suitably elaborated to apply to cases in which supervenience relations are present in the manner applied below)—nothing fancier or richer[8].

The conception just described is very close (perhaps identical) to the understanding of downward causation advocated in Ellis (2016)

> One demonstrates the existence of top-down causation whenever manipulating a higher-level variable can be shown to reliably change lower-level variables

[6] "Causal route" here is intended to contrast with routes or paths corresponding to supervenience relations. Again, we need a somewhat different account of how interventions behave when supervenience relations are present.

[7] The notion of "level" is used in many different and not entirely consistent ways in both science and philosophy. In my view it is doubtful that there is any single characterization of this notion that will fit all these uses. Rather than getting bogged down in trying to provide such a characterization I will rely instead on generally accepted judgments in the scientific literature about particular cases as well as some defeasible criteria. For example, I will assume that variables are often legitimately regarded as at different levels when one is a coarse-graining of the other and that variables used to characterize wholes are often legitimately regarded as at a different level than variables that characterize their parts). For additional discussion, see Woodward, 2020.

[8] I stress this point because, as noted earlier, I think that some of the opposition to the idea that there can be downward causation or causation involving upper-level variables depends on the (often tacit) assumption of a richer or thicker notion of causation and the thought that this sort of causation is not present in relations involving upper-level variables. We should separate the question of whether there are downward causal relations that are causal in the sense of **M** from whether there can be downward causal relations according to some alternative conception of causation. I'm concerned only with the former issue in this essay.

Although this is the basic idea, as I have said, some additional explication is required to specify how it is to be understood in contexts in which variables at different levels are present. To fix our ideas, let us assume that we have two sets of variables *Ui* (for upper) and *Lj* (for lower). Assume that a full specification of the values of the lower-level variables determines the values of the upper-level variables, so that the latter "supervene" on the former. We also assume that different values of the *Ljs* can "realize" the same value of a *U* variable, so that "multiple realization" is present. For reasons described in footnote 8 in most cases of this sort the relationship between the *U*s and the *L*s will *not* be one of identity, either of types or tokens, but will instead amount (in the case in which the *L*s are low- level physical variables) to a version of non-reductive physicalism[9]. In what follows I will assume that such "realization" takes a very specific form: for each upper level variable *Ui* there is a many to one surjective[10] function that maps a number of different values of the *Lj*s into each value of *Ui*. We can think of this function as taking one of two possible forms. One possibility is that a number of different values of the same *Li* variable are mapped into (realize) a single value of a *Ui* variable. As a standard example, think of the values of *Lj* as very high dimensional specifications (profiles) of the possible combinations of kinetic energy that might be assumed by the molecules making up a gas. That is, a single value of *Lj* specifies a possible kinetic energy for molecule 1, a possible kinetic energy for molecule 2 and so on. A different value of *Lj* specifies a different n-tuple of kinetic energies for the individual molecules. A given value of the upper level variable *Ui* (e.g., *Ui* might be temperature T) then can be realized by a very large number of different values of *Li.* Another possibility is that values of several *different* lower-level variables are mapped into the same value of an upper-level variable. For example, the upper-level variable total cholesterol (*TC*) is the sum of the values of two lower-level variables, low density cholesterol (*LDL*) and high density cholesterol (*HDL*). Different combinations of values of *HDL* and *LDL* can realize the same value of *TC*. For the purposes that follow, there are no deep differences between these two possibilities and because it will simplify the exposition I will often just talk about the relationship between an upper level variable *U* and a single realizing variable *L*, assuming that it is obvious how to generalize this to cases in which the realizers of *U* are functions from values of several different *L* variables[11].

[9] A common assumption (which I endorse) is that when the relation between upper and lower-level variables is one of identity there is no particular puzzle about how downward causation and causal relations among upper-level variables are possible: the upper-level variables stand in exactly the same causal relations as the lower-level variables with which they are identical. The issues around downward causation become less trivial when non-identity and multiple realization is assumed

[10] We assume that this function is surjective to capture the standard assumption that every value of each of the *Ui*s is realized by some value (typically many values) of the *Lj*s. For example, any possible value of temperature of a dilute gas is realized by some (typically many) profile(s) of molecular kinetic energies.

[11] I acknowledge that the possibility just described it a very simple one -- I assume it *because* it is simple, because it is one way of making "realization" precise, and because it is in many ways one of least friendly assumptions for the possibility of downward causation. (That is, if downward causation makes sense in such contexts, it is plausible that it will also make sense in contexts in which realization relations that cannot be represented in the simple way I have described.) In this connection I want to explicitly note that there are many other sorts of cases in

In any case, the realization relation is understood as an "unbreakable" constraint relation rather than a causal relationship. It is unbreakable in the sense that the relationship cannot be disrupted by any combination of interventions. For example, although one can manipulate the temperature of a dilute gas (and in doing so will also manipulate the average kinetic energy of its component molecules), one cannot through interventions alter the *relationship* between temperature and average kinetic energy – this is treated as fixed. To anticipate discussion below, when L and U stand in a realization relation the nature of this relationship is such that they are not sufficiently "distinct" to stand in a causal relationship. Thus an upper-level variable U does *not* cause its realizers L[12] and similarly L does not cause U. However, U may cause some other variable lower-level L^* that it distinct from its realizer L. When this is the case, there is downward causation from U to L.

Consider an intervention on an upper-level variable U in a context of the sort just described—e.g., the temperature T of a gas in a container is manipulated by placing it in a heat bath. Different interventions each of which sets T to some value t will be realized by different combinations of values of the lower-level molecular variables Kj on different occasions (for that

science in which inter-level relations are described (at least by philosophers) by means of words like "realization", "constitution" and so on which involve more complex relationships between upper and lower-level variables. For an instructive illustration of some of these complexities in the case of neuronal modeling at different levels, see Herz et al., 2006. In such more complex cases, the variables of the upper level theory may not "line up" in any simple or well-behaved way with the variables of the lower level theory, the mathematics employed at different levels may be quite different (ordinary versus partial differential equations versus black box Bayesian models etc.), and as a result the relations between different levels may be mathematically very complex. Moreover, in many cases, a fully adequate characterization of an upper-level variable will involve reference to what looks like upper-level information as well as information about its lower-level realizers. For example, in the illustration above, I neglected the fact that the notion of temperature of a gas, as usually understood, is only well-defined if the gas is at equilibrium, which is an "upper-level" feature of the whole gas.

One consequence of this complexity is that a good deal of work is often required to connect information at one level to information at different levels—there may not be anything like the simple functional relations I assume above. I will ignore/abstract away from this in what follows. Finally, let me add that the complexity of the relation between upper and lower level variables is one of several reasons why it is often wrong to take this relationship to be one of identity (and, as claimed above, why some form of non-reductive physicalism seems like a more plausible account of this relationship) . An additional consideration is that the most plausible understanding of the notion of identity within the interventionist framework requires a notion of identity between variables and between values of variables. In both cases, a plausible necessary condition is that identical variables (or values) should have the same dimensionality—this of course is violated when there is coarse-graining or dimension reduction of the sort described above.

[12] As noted below, some prominent discussions (e.g., Craver and Bechtel, 2007) proceed on the assumption that if there is such a thing as downward causation it involves an upper-level variable causing its realizer. I agree that upper-level variables do not cause their realizers but argue (Section 4) that this is not what downward causation involves.

matter, the molecular realization of *T* will vary from moment to moment for the same gas)[13]. The experimenter thus controls the value of *T* via the heat bath (that is why it is appropriate to think of the intervention as an intervention *on T*) in the sense that the experimenter possesses a procedure that can reliably and repeatedly impose that temperature. However, the experimenter does not control in this sense which particular values of the molecular kinetic energies that realize that value of *T* – putting the gas in a heat bath is not a procedure that reliably imposes any particular molecular realization of *T=t.* Instead this realization varies from occasion to occasion or over time in a way that is unknown to the experimenter and effectively random from the point of view of what the experimenter can influence[14].

As Ellis suggests, we may think of the values of the variables *Kj* that realize the same value of *T* as in the same equivalence class, yielding a partition of the different values of *Kj* based on this equivalence relation. Of course because of the nature of the realization relationship between *T* and the *Kj*s any intervention that changes the value of *T*, from, say *T=t1* to *T= t2* must at the same time change the values of the lower-level realizing variables *Kj* from values that realize *T= t1* to different values that realize *T=t2*—that is, to a different equivalence class. Contrary to the arguments of a number of philosophers (e.g., Baumgartner, 2010), we thus do *not* build into the notion of an intervention the requirement that an intervention on *U* change the

[13] Thus if one wants to represent such an intervention within a directed graph framework, the appropriate way to do this, as suggested in Woodward, 2015, is by means of a single intervention *I* that sets *T=t* and at the same time "selects" some value from the equivalence class of lower level realizers of *T=t.* There are not two different interventions, one that sets *T=t* and distinct from this a separate, independent intervention the intervention that sets the value of the lower level realizer of *T=t.* It is a also mistake to represent such an intervention as a common cause of both *T* and the realizing variable or variables *Kj*, as, for example, Baumgartner 2018 does —that is to represent the intervention as $Kj \leftarrow I \rightarrow T$. It follows from standard assumptions made about causal representation in directed graphs (including, for example, the condition of independent fixability (**IF)** described below), that such a common cause representation would only be appropriate if it were possible to intervene to carry out independent interventions on *T* and the *Kj*, changing each independently of the other. The realization relationship between *T* and *Kj* rules out this possibility. This is not a pedantic point because the common cause representation is used by Baumgartner and others to motivate the claim that one needs to control for *Kj* in assessing the causal effect of *T* and hence immediately to a causal exclusion argument according to which *T* is causally inert—again see below.

[14] In some cases, including the case of temperature discussed above, it may be reasonable to assume that for each value of the upper level variable, there is a single stable probability distribution over the values of the lower-level realizers of the upper-level variable that applies whenever there is an intervention on the upper-level variable. However, in many other cases, this will not be a plausible assumption and I do not adopt it in what follows. The requirement described below that the realizers of each value of the upper-level variable have a uniform effect on the effect variable of interest amounts to the assumption that such uniformity holds for *all* probability distributions over the values of the realizing variables. However, there are various ways of relaxing this requirement, one of which is simply to require that uniformity hold only for all "reasonable" probability distributions, where "reasonable" might mean, e.g., "absolutely continuous with respect to the Lebesgue measure." Other possibilities for relaxing the uniformity requirement are described below.

value of *U* while leaving values of the lower variables *Lj* that realize that value of *U* unchanged. Such interventions are impossible (because the realization relation is an unbreakable constraint); adopting such a requirement would have the consequence that interventions on upper-level variables are never possible and would render this notion useless for purposes of understanding upper-level causation.[15] (Recall that **M** requires that for a variable *X* to have a causal effect, interventions on *X* must be possible.)

In order to apply **M** to contexts in which different levels are present, we must also impose the following requirement (called *realization independence* in Woodward, 2008): when values of *U* are realized by a number of different values of *Ls*, 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*. In other words, an intervention that sets *U=u*, must result in the same response for *Y* (*Y=y*), regardless of how *U=u* is realized at the lower level[16].

Here again Ellis imposes a closely related requirement: "the same top level state must lead to the same top level outcome, independent of which lower level state instantiates the higher level state". (2016, 121)

The effect of this requirement of realization independence is to exclude so -called "ambiguous manipulations" (Spirtes and Scheines, 2004) in which the result of setting *U=u* on some second variable *Y* depends on how *U=u* is realized. To illustrate, suppose that the lower level variables are *HDL* and *LDL* (as discussed above) with *HDL* having a favorable effect on heart health and *LDL* an unfavorable effect. The upper-level variable *TC* (total cholesterol) which is the arithmetic sum of *HDL* and *LDL* will fail the realization independence requirement with respect to heart health since the impact of *TC =tc* on heart health will depend upon the particular combination of values of *HDL* and *LDL* that realize *TC= tc*. One way of motivating this requirement is to note that it is needed for the effect on *Y* of an intervention on *U* to be well-

[15] More technically, in contexts in which a realization relation between *U* and *L (*or some set of *Ls)* is present, the requirement in Woodward, 2003 that an intervention *I* on *U* with respect to a second variable *Y* not affect *Y* via variables on paths that do not go through *U* ("off path variables") should be understood in such a way that the variable *L* which realizes *U* is not treated as such an "off-path" variable. This corresponds to the idea that *L*s should not be treated as potential confounders for the *U*→ *Y* relationship which we have to "control for" to see the effect of *U* on *Y*. Some additional justification for this (which seems to me a common sense requirement) is provided in Woodward, 2015 and also below (Section **7**).

[16] As several writers note (e.g. Butterfield, 2012, Rubenstein et al, 2017, we can think of this uniformity requirement as amounting to a kind of coherence or consistency requirement between the causal relations involving upper and lower-level variables. Given some natural additional assumptions (described in Rubenstein, et al, 2017), it is equivalent to the following "commutivity" requirement: Suppose F describes the lower-level functional relationship between *L1* and *L2*, g1 describes the realizing relation that maps *L1* to *U1,* g2 the realizing relation between *L2* and *U2* and H describes the upper-level causal relation between *U1* and *U2.* Then for *U1* to cause *U2* and for consistency across levels, the result of beginning with some value of *L1*, applying F to it to yield *L2* and then coarse graining *L2* via *g2* to yield some value for *U2 = u2* should be the same as beginning with the same value of *L1*, coarsening it via *g1* to yield a value of *U1*, and applying *H* to *U1*—this should yield the same value of *U2= u2* as before.

defined: this requires, as Ellis, says, that there be a "regular" or "same" response of *Y* to the intervention on *U*. This implies that to the extent that we are interested in effects on heart health, *TC* is not a "good" upper-level variable—not a good candidate for an upper-level cause. It should be replaced by variables that have unambiguous (or at least less ambiguous) effects on heart health.

Note that the requirement of realization independence, like the notion of an intervention itself is always defined relative to a candidate effect variable. It common for an intervention on *U* that satisfies the realization independence with respect to *Y* to fail to satisfy this requirement with respect to some distinct variable *Y**.

The conditions described are, I believe, necessary for downward causation but I do not claim they are jointly sufficient[17]. However, I believe it is plausible that whatever additional conditions may be required for sufficiency are satisfied for the examples of downward causation I will discuss below—or so I will assume.

[17] Why might one think that the conditions described above are not sufficient? My doubts arise from the following consideration. It looks as though an upper level variables *U1* might meet those conditions and yet be (at least from our perspective) highly gerry-mandered, non-compactly distributed and difficult to recognize, measure or manipulate. Consider tosses of a fair coin. We might form the equivalence class of all those initial conditions of the coin and the tossing apparatus that lead to heads – take all these to have the value *h*-- and the equivalence class of those conditions leading to tails (these have the value *t*). We might then form the upper level variable *C* which takes the values *h* and *t*. By construction the values of *C* have a uniform effect on the final position of the coin. But whether or not *C* is an "in principle" legitimate upper -level variable or candidate cause, it is certainly not a *useful* variable, assuming that we have no way of telling, apart from the final position of the coin, which value of *C* is realized in any particular toss, no way of manipulating *C* and so on.

A natural thought which is suggested in passing by Ellis, is that at least in many cases in which we find it natural to talk of upper level or top-down causation, we expect some additional condition to be satisfied that excludes cases of the sort just described: we want the candidate upper-level variable to correspond to something we can measure with relatively macroscopic (upper-level) measurement procedures and manipulate by means of macroscopic interventions, where we require such interventions to have coordinated or orderly effects on lower-level variables. This expectation is fairly well satisfied in connection with thermodynamic variables—we have straightforward procedures for measuring and manipulating these – e.g., by putting the gas in a heat bath or by compressing it with a piston. When we do this we think of ourselves as imposing a co-ordinated change in the behavior of the constituents of the gas. This goes along with the more general thought that talk of upper-level causation seems most appropriate when there is a kind of order or co-ordination or coherence in the behavior of the lower-level constituents that realize the upper-level variables, with the loss of such order corresponding to cases in which causation resides more exclusively at lower levels, as the example involving energy cascades in Section **3** illustrates. There are connections here with the distinction between work and heat.

Finally, there are two other conditions on causation in general (and not just downward causation) that will play a role in my discussion. The first is the requirement that the relata that figure in causal relations must be *variables* (which, I remind the reader, is my shorthand for what in the world is represented by variables). This is also a requirement that Elllis imposes, as reflected in the passages quoted above. Variables represent quantities or magnitudes (e.g., mass, charge, income) or, as a limiting case, whether some property is present or absent (represented by a binary variable taking the values 1 and 0.) As this suggests, one mark of a variable is that must be capable of taking at least two distinct values. This requirement might seem trivial but as we shall see, neglect of it (or failure to specify just what the relevant variables are) undermines some well - known criticisms of downward causation.

A second generally accepted requirement on causation is that variables standing in causal relationships must be "distinct"—the intent here is to rule out cases in which variables stand in logical, conceptual or state-space relationships that exclude causation. For example (Lewis, 2000), although whether or not I say "hello loudly" depends in some sense on whether or not I say "hello", this dependence is not causal dependence. I will provide a characterization of the kind of distinctness that is necessary for causation below—condition **IF**, Section **5**. The relevance of this consideration to our discussion is that critics of downward causation frequently claim that this involves wholes acting downward on their parts and that wholes and parts are not sufficiently distinct to stand in causal relationships. (See e. g. Craver and Bechtel, 2007) I agree that at least in many cases wholes and their parts are not sufficiently distinct to stand in causal relationships but, as argued below, in other respects this criticism misfires. Scientifically plausible examples of downward causation do not involve wholes acting on parts but rather involve *variables* (as all causal relations do) and these need not stand in part/whole relationships, even when entities of which they are predicated do.

3. Some Examples

Recent papers and books by Ellis and co-authors and by others such as Denis Noble provide many prima-facie plausible examples of downward causation. (See also Clark and Lancaster, 2017.) Here are a few such examples, with some additions of my own. (Again, in saying that these are "prima-facie plausible" examples I do not mean that I'm going to simply assume that these are genuine cases of downward causation. Rather, following the methodology outlined earlier, these are the kinds of cases that count as downward causation if any cases do and hence the kinds of cases on which we need to focus.)

3.1) The use of mean field theories in which the combined action of many atoms on a single atom is represented by means of an effective potential V rather than by means of a representation of each individual atom and their interaction. Intuitively, V is a higher level than the atom on which it acts. (Ellis, 2016, Clark and Lancaster, 2017).

3. 2) The influence of environmental variables including social relations between animals on gene expression as when manipulating the position of a monkey within a status hierarchy changes gene expression which controls serotonin levels within individual monkeys. Here position within a social hierarchy is thought of (perhaps on the basis of compositional considerations) at a higher level than gene expression.

3.3) A red hot sword is plunged into cold water and this alters the meso -level structure of the steel in the sword— cracks, dislocations, and grains that it contains. The treatment of the sword—heating and cooling—is at a higher level than these mesoscopic changes[18] and the former downward causes the latter. (Example due to Bob Batterman.)

3. 4) Energy cascades. When a fluid is stirred in such a way that it exhibits large-scale turbulent motion this motion is gradually transferred to motion at smaller scales– from large scale eddies to much smaller scale eddies. The large-scale motion may be on the scale of many meters, the small-scale motions on the scale of a millimeter where they are eventually dissipated as heat. Viscosity related effects dominate at this smaller scale but are less important at larger scales. The stirring is an upper-level cause of the subsequent behavior of the fluid. (Example due to Mark Wilson.)

3.5) According to the Hodgkin- Huxley (HH) model, a neuron generating an action potential may be represented by a circuit in parallel, in which there is a potential difference V across the neuronal membrane which functions as a capacitor. Embedded in the membrane are various sodium and potassium ion channels with time and voltage dependent conductances g_{Na}, g_K which influence ionic currents through the membrane. V causally influences these conductances and currents which seem intuitively at a lower level than V. (Example discussed by Denis Noble, 2006.)

In each of these cases the conditions for an intervention on the upper-level variable seem to be satisfied[19]. First, the manipulations of the upper-level variables are not confounded by other variables that might affect the dependent variable independently of the intervention in a way that undermines the reliability of causal inferences. (Some writers -- e.g., Baumgartner, 2010—hold that all manipulations of upper-level variables are "confounded" by their lower level realizers but this is a tendentious and unmotivated notion of confounding—see Section 7.) Second, the upper-level variables are multiply realized but it is plausible that their effects on the dependent variables are realization independent in the sense described in Section **2**. For example, there are a variety of different ways of intervening on the mean field to set it to a particular value (with these corresponding to different arrangements of the many atoms making up this field) but as long as the value of the mean field is the same, the effect on the individual atom will be the same. In the case of **3.5**, interventions on the membrane potential can be carried out by means of a voltage clamp (the device actually employed by Hodgkin and Huxley in carrying out their original experiment) which exogenously imposes a stable potential difference across the membrane. A particular value for this potential difference can be realized at a lower level by various combinations of charge carrying individual atoms and molecules in the membrane but to

[18] The heating and cooling affect the whole sword, not just components of it.

[19] Recall that according to **M** for causal claims to be true the interventionist account does not require that interventions actually occur but rather the truth of the appropriate counterfactuals describing what would happen if interventions were to occur. However, in the examples described above, interventions are actually carried out to demonstrate downward causation—for example, position of a monkey within a status hierarchy is manipulated and the effect on its serotonin level observed.

the extent that the HH model is empirically correct, these different realizations will have the same uniform impact on lower-level variables such as the channel conductances.

Although examples of the sort just described appear to be prima-facie plausible examples of downward causation, a number of scientists and philosophers have advanced objections to this concept. In the next several sections (**4- 7**) I review and respond to several of these objections.

4. Wholes and Parts.

A very common criticism of the idea of downward causation is that this requires that "wholes" act downward on their "parts" and that the relation between a whole and its parts cannot be a causal relation of any kind. Two reasons (e.g., Craver and Bechtel, 2007) cited in support of this last claim are that (i) wholes and parts are not sufficiently distinct to stand in causal relations and (ii) the relation between wholes and parts is "synchronous" while causal relationships are always "diachronic", where this is understood as meaning that effects must occur temporally after their causes. For example, Craver and Bechtel, 2007 consider, as a putative example of top-down causation, the claim that the overall process of visual signal transduction (from light falling on the retina to visual object recognition) causes changes in the components or parts of the transduction process such as rod depolarization—i.e., that this whole temporally extended process causes the occurrence of its temporal components. They object that because rod depolarization is part of the overall transduction process, the latter cannot cause the former. More generally, they think of claims of downward causation as claims that the overall state or activity of a mechanism has instantaneous or synchronic causal effects on components of the mechanism—a notion that they find objectionable.

A basic problem with this line of argument is that plausible cases of downward causation, including the examples described in Section **3**, do not take this whole to part form. One reason for this is that parts and wholes are (at least on the most natural interpretation of these notions) *things or thing-like* (where included in the latter category are temporally extended processes or, as some philosophers call them, "activities"). By contrast, as emphasized above, causal claims relate variables and at least in many cases these variables do not stand in part/ whole or containment or constitutive relationships. This is so even if it is true that the things of which these variables are predicated stand in part/whole relationships. For example, in the case of the HH model, the putative top-down cause is not the whole process of the generation of the action potential. Rather the top-down cause is changes in the membrane potential *V*, a variable (more pedantically a magnitude represented by a variable), and among its effects are changes in the voltage-gated channel conductances, represented by the variables g_{na}, g_k . The ion channels, the conductances of which are described by g_{na}, g_k are indeed parts of neuronal membrane but it does not follow (indeed it is unclear what it would mean to say) that the conductances are themselves parts of the membrane potential difference[20]. More importantly, even if we think that

[20] Craver (2007) does provide a test for whether some activity or behavior is a "part" of another. This appeals to what Craver calls mutual manipulability (MM): when X and S are related as part and whole and F is an behavior of X and J a behavior of S, then F is a constituent or part of J iff

(i) there is an intervention on X's F-ing with respect to S's J-ing that changes S's J-ing;

(ii) there is intervention on S's J-ing with respect to X's F-ing that changes X's F-ing (Craver, there is a way of making sense of this parthood claim, it does not follow, for reasons described below, that the membrane potential and conductances fail to be distinct in a way that precludes their standing in a causal relationship.

Similarly, when a heated sword is plunged into cold water, it is true that the meso-structures affected are parts of the sword, but the relevant causal claim is not that the sword or its overall state causes these meso- structures to change instantaneously or that the temporally extended process consisting of plunging the sword followed by lower-level structural changes causes the temporal part consisting of the latter changes. Instead, the top-down cause in this case is the act of plunging the hot sword into the cold water which might be represented by a binary variable *P* which takes the values 1 or 0 depending on whether the sword is or is not plunged. Again, the meso-structure of the sword is not plausibly regarded as "part" of the variable *P*.

Similarly, monkey 1 is (let us suppose) a "part" of the monkey band, and monkey 1's serotonin level is part of monkey 1 but the putative top down cause (and what is experimentally manipulated) is the hierarchical structure of the band and the putative effect (monkey 1's serotonin level) is not (at least in any obvious sense) part of that.

These distinctions (between things or processes which have parts and variables which at least in the cases under discussion do not stand in part/whole relations) would not matter if whenever P is a part of whole W, variables predicated of P and W fail to be distinct in a way (or have some other property) that precludes their standing in causal relationships. However, as I shall now argue, this is not the case: as the examples in **3.5** illustrate, even if P is a part of W, it does not follow that variables predicated of P and W cannot stand in causal relationships.

5. Independent Fixability.

The following condition is commonly assumed, often only implicitly, in the causal modeling literature for when variables are sufficiently distinct to stand in causal relationships. I call it **IF** (for Independent Fixability) since it embodies the idea that variables are distinct if all of their values are independently fixable via interventions:

> (**IF**) Variables in set **S** are distinct in a way that permits their standing in causal relationships if and only it is "possible" to intervene on each variable independently, holding it fixed at each of its possible values (for the units or systems those values

2007 , p. 153).

This is a test for whether activities/ behaviors rather than variables are parts of others, but putting this aside, MM is inadequate because it fails to distinguish genuine parthood relations from cyclic causal relations. For example if having a certain potential is a behavior then both (i) and (ii) are satisfied with respect to the relations between the potential and the behavior of the channel conductances g_{na}, g_k. However, both the *V* to g_{na}, g_k relation and the g_{na}, g_k to *V* relations are causal rather than whole/part relations. (See Section **6** for remarks defending the claim that causal relations can be cyclic.) As argued in Section **5**, the feature of a part/whole relation that precludes causation is a failure of independent fixability. This is present in Lewis example of the relation between saying "hello" and saying hello "loudly" but not in the case of the relation between V and g_{na}, g_k

> characterize) while intervening to hold the other variables to each of their other possible values. In other words, all possible combinations of values of different variables in the set must be "compossible"[21]. Here "possible" includes settings of values of variables that are possible in terms of the assumed logical, mathematical, or semantic relations among the variables as well as certain structural or space-state relationships.

As an illustration, consider an example from Lewis (1986) concerning of the relationship between N's saying "hello" and saying "hello" loudly. Let *H* be a variable that takes the values 0 or 1, depending on whether or not N says "hello". Let *L* be a variable that takes values 0, 1 or 2 depending on whether N does not say hello", says "hello" but not loudly, or says "hello" loudly. Then certain combinations of these variables such as *H*=0 and *L*=2 are impossible for conceptual reasons and **IF** is violated. Thus, as Lewis claims, the relationship between *H* and *L* is not a causal relationship. As another illustration, the variables in {*HDL, LDL* and *TC*} are conceptually connected and fail the independent fixability condition: Given, e.g., values for *HDL* and *LDL*, there are values for *TC* that are ruled out for mathematical or conceptual reasons, since *TC* is defined as the sum of *HDL* and *LDL*. This is reflected in the fact that it would be misguided to claim that *HDL* and *LDL* cause *TC* For similar reasons, **IF** is violated for upper and lower level variables that stand in realization relations—a variable (with n-tuples as values) representing the kinetic energies of all of the individual molecules in a gas cannot cause its temperature (or conversely.)

Fortunately to apply **IF** to the putative examples of downward causation in **3.5,** we do not need to make problematic judgments about logical or conceptual possibility. In each case, the possibility of independent fixability is shown by the fact that experiments have actually been performed (or might readily be performed) that set the values of the variables claimed to be causally related independently of each other. For example, in the experiments which provided the basis for the HH model, the newly invented voltage clamp allowed the experimenters to set the value of *V* exogenously in a way that was independent of the channel conductances. Similarly, the channel conductances can be manipulated independently of *V* by molecular agents. In the case of **3.2,** the status position of a monkey can be changed by placing him in a new band and observing whether there are changes in his serotonin level. Also the serotonin level of the monkey can be manipulated independently by pharmacological means. These possible experiments reflect the fact that the variables in the relationships **3.1-3.5** do not seem to be logically or conceptually connected in a way that precludes their standing in causal relationships.

Another concern expressed by Craver and Bechtel, 2007, as well as other writers, is that putative relationships of downward causation are synchronic while legitimate causal relations are diachronic, with the cause temporally preceding the effect. Again this concern seems to derive from the mistaken assumption that downward causal relationships are whole to part relationships, where these are understood as obtaining instantaneously or at single moment. In fact the general claim causes must always temporally precede their effects is far from obviously correct but it is not necessary to argue for this here, since the examples **3.1- 3.5** all seem to

[21] On the other hand, different values of the same variable are not compossible for the same object or system, in the sense that such different values cannot hold for the same object —e.g., the same object cannot have a mass of both 5 and 10 kg. Of course different objects can have different masses, and the velocity of any object can be set independently of its mass.

involve diachronic causation, although this fact may not be represented in the way those relationships are modeled or described. For example, as an empirical matter, there is presumably a very short temporal delay between the momentary value of the membrane potential or its time derivative and the response of the ion channels, although this fact is not represented in the HH model, since it does not matter to the effects that the model is intended to explain. Similarly the response of the monkey's serotonin levels to a change in status presumably also does not occur instantaneously but rather takes time. If, like Craver and Bechtel, one understands part/whole relationships as those that obtain at a given instant, such relations are indeed "synchronous" but this is just further reason to think that examples involving such relations are very different from the relationships described in **3.1-3.5** and not plausible candidates for causal relations of any kind.[22]

6. Cycles.

Another concern about downward causation that appears in Craver and Bechtel, 2007 is this: it appears that countenancing downward causation in a system leads, in many cases, to countenancing causal cycles in that system, in the sense that at some level of representation we have *X* causing *Y* which in turn causes *X* (perhaps via some intermediate variables). Craver and Bechtel claim that such cycles are problematic— because (among other considerations) they are inconsistent with the "asymmetry" of casual relationships[23]. The claim that in a number of cases[24] systems in which downward causation is present will also be systems in which cycles are

[22] It is worth noting that the examples Craver and Bechtel discuss appear to be ones they have made up—they don't cite anyone in the scientific literature who treats their examples as cases of downward causation.

[23] An anonymous referee suggests this may be a misunderstanding on my part since in subsequent papers (e.g., 2017) Bechtel does discuss causal cycles. But in their (2007) Craver and Bechtel are unambiguous that they think that downward causation as they conceive it is problematic because it seems to involve causal cycles:

> …the possibility of bottom-up and top-down influence 'propagated' simultaneously across levels results in problematic causal circles. For example, one might believe that if an object, X, has its causal powers in virtue of possessing a property, P, then if X is to exercise its powers at time t, X must possess P at t. And one might believe further that if something causes X to acquire P at t, then x does not already possess P at t until that something has acted. If X's acquiring P at t is a cause of S's having w at t, and S's having w at t is a cause of X's having P at t, then it appears that X's acquiring P at t cannot cause S to have w until S's having w causes X to acquire P. In that case, it is little wonder that talk of interlevel causation strikes us as mysterious. (552-3).

I will not speculate about how to reconcile these remarks with Bechtel's later remarks regarding causal cycles in mechanisms.

[24] This is not true for all examples involving downward causation as shown by **3.3** and **3.5**.

present seems correct. For example, as we have noted, in the case of the HH model, the membrane potential causally influences the channel conductances but it is also the case that those conductances, by influencing ion flow, in turn influence the membrane potential. Similarly, although status position influences serotonin levels, it is also the case that serotonin levels influence status, as is shown when the former is exogenously manipulated.

Causal representations involving cycles raise a number of subtle interpretive issues that I lack the space (and competence) to address. However let me make the following brief points:

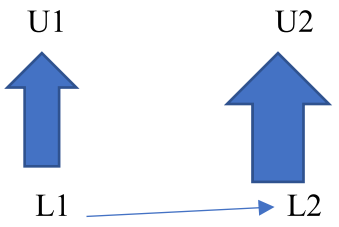
6.1. The presence of causal cycles is a real feature of many biological systems and for obvious reasons—such cycles are an unavoidable part of feedback and control mechanisms that are ubiquitous in such systems and are necessary for restoring systems to an equilibrium from which they may have departed, avoiding runaway behavior etc. Cycles are also a common feature of many social and economic systems. We don't want conditions on causation that have the consequence that such cycles are impossible.

6.2. One possible strategy for avoiding cycles is to distinguish variables by assigning them different temporal indices: the membrane potential at time *t* (V_t) causally influences the conductances at time *t*+ *d*, which in turn influence the membrane potential at time *t* + *2d* represented by a distinct (V_{t+2d}) and so on. I will not pursue the question of whether this strategy is always appropriate but it is one way of replacing cycles with non-cyclic systems.

6.3. Representations involving causal cycles (that is, that do not employ the indexing strategy described under **6.2**) are common in the causal modeling literature (these are so-called non-recursive models) and in disciplines like economics. It is not obvious that there is anything incoherent (or inconsistent with the "nature" of causation) in the use of such models, even if we think that underlying them at some finer-grained level of analysis is an acyclic model. In thinking about representations with cycles, we should distinguish the issue of whether they postulate relations that have a "direction" from whether causal cycles are possible; directionality is arguably a feature of any causal relationship, in the sense that we haven't specified the relationship until we have specified a direction and that $X \rightarrow Y$ is a different relation from $Y \rightarrow X$. However, this does not preclude cycles. In a causal graph in which $X \rightarrow Y$ and $Y \rightarrow X$ appears, the graph is directed (rather than undirected, as would be the case if we had instead written *X*--*Y*) reflecting the fact that there is a causal relation is from *X* to *Y* as well as from *Y* to *X* but a cycle is present. In other words, we need to distinguish directedness from acyclicity: there can be directed cyclic graphs as well as directed acyclic graphs. A simple interpretation for such a directed cyclic graph (which will fit some applications but perhaps not all) is this: There is an intervention on *X* that will change *Y* and an intervention on *Y* that will change *X*. This seems to fit the examples in section **3** in which there are apparent causal cycles—intervening on status changes serotonin levels and intervening on serotonin levels changes status and so on. There does not seem to be anything incoherent about such an interpretation.

7. Causal Exclusion.

Another objection to the notion of downward causation appeals to causal exclusion arguments. Suppose, as before, that *U1* and *U2* are upper-level variables and *L1* and *L2* are lower-level variables with *L1* realizing *U1* and L2 realizing *U2*. *L1* causes *L2*. An iconic diagram due to Kim (e.g., 2005), represents these the realization relations by means of a thick vertical arrow and the causal relation by means of a thin horizontal arrow:



The question is then whether there can be (whether it makes sense to suppose that there are) other causal relationships in this structure; for example, between *U1* and *U2* or between *U1* and *L1* (the latter being a case of "downward causation"). According to the causal exclusion argument the answer to this question is "no" and thus downward causation (as well as upper-level causation from *U1* to *U2*) is impossible. A number of different but closely related arguments are invoked in support of this conclusion. Here is one: (i) Assume for simplicity that the lower-level causal relation is deterministic. Because of the realization relation, a change in the value of *U1* must involve a change in the value of *L1*. Suppose that under this change there is an accompanying change in the value of *L2*. (If there are no changes in the value of *L2* that accompanies changes in the value of *U1/L1*, then *U1* doesn't cause *L2* and it also does not cause *U2*.) This change in the value of *L2* under a change in *L1* shows that *L1* causes *L2*. Moreover (according to this version of the exclusion argument) this change in *L2* is" entirely due" to the change in *L1*, so that there is no "causal work left over for *U1* to do" when it comes to *L2* (or *U2*). In other words, *U1* appears to be causally inert with respect to *L2* once the role of *L1* is taken into account.

A related argument (ii) claims that countenancing downward causation from *U1* to *L2* commits us to an implausible and unnecessary claim about causal overdetermination: if downward causation was present we would have both *U1* and *L1* causing *L2*. Not only does this seem "counterintuitive" according to critics, postulating such overdetermination seems unnecessary, since as we have seen, any effect on *L2* seems to be fully accounted for by *L1* alone—postulating a causal influence from *U1* to *L2* is (it is claimed) superfluous or redundant.

Finally, (iii) suppose we want to determine whether *U1* has a causal impact on *L2*. To do this, we must, according to advocates of the exclusion argument, "control for" the causal influence of other causes of *L2* besides *U1*—it is only if *U1* influences *L2* holding fixed (that is conditioning on) or accounting for the influence of these other causes, that we can conclude that *U1* causes *L2*. But among the "other causes" of *L2* is *L1* and once we control for the influence of *L1* on *L2*, we see that *U1* has no further or additional effect on *L2* – indeed, given the value of *L1*, any further variation in *U1* (which might be responsible for any additional effect of *U1*) is impossible (cf. Baumgartner, 2010).

I have discussed these arguments elsewhere (e.g., Woodward, 2015). Here I will be brief: on my view, they rest on misunderstandings about how to think about causal relationships when non-causal determination relationships (like the realization relation between *L1* and *U1*) are also present. Let me begin with version (iii) of the exclusion argument, since the mistake here is perhaps most obvious. Suppose that we have a structure S* (figure 2, with the thin arrows representing causal relations) in which, in contrast to the structure S in Figure 1, *L1 causes* (and thus does not realize) *U1* and in addition *L1* causes *L2* which causes *U2*. In this case *U1* and *L2* will be correlated as will *U1* and *U2*. In cases of this sort in determining whether *U1* causes *L2* (or *U2*), it is indeed entirely appropriate to control for the influence of *L1* on *L2*: *U1* causes *L2* only if , taking into account the influence of *L1* on *L2*, *U1* has an additional independent effect on *L2*. If the correct structure is what is represented in the diagram, when one controls for *L1*, *U1* will not be correlated with *L2* or *U2*, showing the absence of a causal connection.

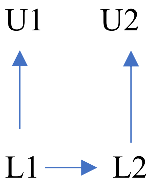


Figure 2: Causal Structure S*

The basic mistake made by defenders of the exclusion argument is to suppose that we are entitled to reason in the same way when the causal relationship between *L1* and *U1* in Figure 2 is replaced with a non-causal determination relation like realization as in Figure 1, so that the same test for whether *U1* causally influences *L2* is appropriate in both cases. In fact the two situations, S and S* are fundamentally disanalogous. For one thing, in situation S* the relevant counterfactual has a possibly true antecedent, indeed one that may be experimentally realizable: one holds fixed the value of *L1*, manipulates *U1* independently (it follows from **IF** that this will be possible in principle if the relationship between *L1* and *U1* is causal or correlational, as in S*, and not one of non-causal dependence) and then sees whether there is any uniform change in the value of *L2*—this is the appropriate criterion for whether *U1* has a causal influence on *L2*. (Parallel remarks apply to whether *U1* causes *U2*.) If it turns out that there is no regular association between *U1* and *L2* in this circumstance, this does indeed allow one to conclude that *U1* is causally inert with respect to *L2*. But in situation S the corresponding counterfactual has, by hypothesis, an "impossible" antecedent: because of the nature of the realization relation, it is impossible to hold fixed the value of *L1* while performing interventions that change the value of *U1* and seeing what changes may be associated with these. This is an indication that the use of this counterfactual is the wrong test or criterion for whether *U1* has a causal influence on *L2*. Put differently, in situation S* the conclusion that *U1* is causally inert follows if it is *possible* to vary *U1* while *P1* is fixed and there is no corresponding change in *L2*. In situation S , the claim of the exclusion argument is, in effect, that the causal inertness of *U1* with respect to *L2* follows from the *impossibility* of varying *U1* while *L1* is held fixed[25]. This relies on a condition for causal inertness that is completely different from the condition employed in connection with S*: a

[25] Recall that the interventionist condition for causation (**M**) requires that there exist possible interventions on X such that…. . Thus cases involving impossible interventions correspond to false causal claims.

condition that is not supported by (is not a reasonable extension of) ordinary considerations about what it is appropriate to control for when only ordinary causal relationships and no relations of non-causal determination are present[26]. Note also that the characterization of interventions on upper-level variables when they have lower- level realizers in Section 2 avoids the problem just described concerning inappropriate control because according to that characterization an intervention that changes an upper-level variable at the same time is accompanied by some change in the lower-level realizer of that variable. That is, when we intervene on *U1* the upshot of that very intervention is also some change in *L1* that, whatever it may be, is consistent with the change in *U1*. This ensures that, assuming *U1* has a uniform or realization independent effect on some second variable *Y* (upper or lower level), this must be consistent with any change in *Y* due to the change in *L1.* In other words when we intervene on *U1* we just let *L1* change in whatever way it does consistent with the intervention and this gives us the effect if any on *Y*.

Another, related way of bringing out why it is inappropriate to control for *L1* in assessing whether *U1* causes *L1* or *U2* in cases in which *L1* and *L2* are realizers of *L1* and *L2* appeals to the underlying rationale for such control – what we are trying to accomplish when we control for potential confounders. Suppose, to take a concrete example, that we are interested in whether administration *A* of a drug X causes recovery *R* from an illness. To answer this question it is not enough to observe whether there is a correlation between *A* and *R*. It might be the case that the drug was preferentially given to those with very strong immune systems (*S*) and that this has an effect on recovery that is independent of the drug. To show that *A* causes *R*, we need to rule out such possibilities. We can do so either by means of a randomized controlled experiment in which the possible confounding influence of *S* is eliminated by the experimental design or, if the study is observational, by measuring *S* and conditionalizing on (controlling for) it. One obvious motivation for doing this is that if the correlation between *A* and *R* is entirely due to *S*, then when we give the drug to those without strong immune systems this correlation will disappear. Thus we will be misled if we attempt to use the drug to promote recovery in a population with this different value of *S*. Note that this is a real worry because it is entirely possible to give someone the drug without that person having a strong immune system.

Call the randomized experiment described above *experiment one* and suppose that when we do it we do get convincing evidence that *A* causes *R*. Now contrast this with the following possibility which I will call *scenario two*. Professor Exclusion observes that drug X has microstructure Q and objects to experiment one on the following grounds: in assessing the possible causal influence of *A*, the experimenters failed to control for Q which is also a cause of recovery (or at least of whatever microlevel facts "underlie" recovery.) Professor Exclusion argues that it is plausible that *A* has no causal influence on R "over and above" the influence of Q, and concludes from this that *A* does not cause *R*. I think it is obvious that Professor Exclusion's worry is completely different from the worry that *S* might be a confounding

[26] Another way of putting this point is that a graph like that in Figure 1, in which realization relations are represented, is *not* a causal graph (that is a graph in which all arrows represent causal relationships) in the sense in which such graphs are understood in, e.g., Pearl, 2000, Spirtes, Glymour and Scheines, 2000 and Woodward, 2003. Instead it is a "mixed" graph in which both causal relations and non-causal (e.g. supervenience relations) are present. Such mixed graphs require different rules for characterizing the effects of interventions and what needs to be controlled for in order to "see" causal relationships.

influence that is addressed in experiment one. First, unlike experiment one that addresses the possible confounding role of *S*, there is no possible experiment that consists of controlling for Q while varying *A.* Second, as noted above if the association between *A* and R in experiment one is entirely due to *S*, then that association will disappear when the drug is given to those with weak immune systems—that is, when there is a change in the value of the confounding variable *S*. In contrast nothing like this is possible under scenario two. The relationship between drug X and its microstructure Q is unbreakable—you don't have to worry that, you might be in a situation in which although you administer X, its alleged confounder Q is absent. In other words the kind of concern about the consequences of confounding which is addressed in the first experiment just isn't a concern in the second scenario. This suggests in turn that there is no obvious motivation for treating Q as a potential confounder that needs to be controlled for. To be sure, from Professor Exclusion's perspective when you fail to conclude that *A* is causally inert you make a mistake, but the point is that this alleged mistake has no further consequences you should care about—it doesn't imply that you will be mistaken about which relationships support manipulation and control, what will happen when you manipulate *A* and so on. On the contrary, from a functionalist perspective you make a mistake when you control for Q since this mistakenly leads you to conclude that *A* does not cause *R*, hence that manipulating *A* is not a way of changing *R*, when, supposing that interventions are understood along the lines described above, there is a manipulation-supporting relationship between *A* and *R*.

To this we may add the following consideration: in the argument immediately above I focused on the use of an exclusion argument to criticize downward causation. But of course if the considerations in the various versions of the exclusion argument are cogent at all, they appear to apply not just to downward causation claims but to all claims that attribute causal efficacy to upper-level variables as long as these variables are not identical with lower level variables-- that is, all claims according to which upper level causes cause upper-level effects (at the same level) turn out to be false to be false as well, under the assumption of non-reductive physicalism. Needless to say, the conclusion that there is no causation at all involving upper-level variables is a difficult one to swallow—it is reasonable to suspect that something may be wrong with premises that lead to this conclusion, which is what I have suggested[27].

What about the overdetermination argument? Again this seems to trade on a misleading analogy (or assumption of similarity between) ordinary cases of overdetermination in which the variables involved do not stand in any non-causal determination relations) and a (very different) kind of "overdetermination" which may occur when such non-causal determination relations are

[27] Put slightly differently, the defender of the exclusion argument seems to claim that built into our notion of causation is a requirement to control for lower-level realizing variables (or at least that we ought to adopt a notion of causation that has this feature). This in turn has the consequence that, under the assumption of non- reductive physicalism, upper-level variables are always causally inert, thus depriving the notion of causation of much of its usefulness since it follows that there are no true upper-level claims, completely independently of any empirical investigation. An obvious question is why we would have developed (and continue to use) a notion of causation with this perverse feature One obvious response is that our notion either does not have this feature. Alternatively, one might think that if it does, it should replaced with a notion that does not have this feature. In fact, recent psychological experiments (Blanchard et al., forthcoming) seem to show that ordinary people do not employ notions of causation that behave in accord with exclusionist assumptions.

present. Consider an ordinary case of overdetermination in which two riflemen both simultaneous shoot ($S1$, $S2$) a victim through the heart with each shot being causally sufficient for death (D). In such a case we may assume that the following counterfactuals are true[28]:

7.1. If $S1$ had not occurred but $S2$ had occurred, D would have occurred

7.2. If $S2$ had not occurred but $S1$ had occurred, D would have occurred

These counterfactuals capture an important part of what makes this an ordinary case of overdetermination. Note that the antecedents of both counterfactuals are possible—one of the riflemen might have decided not to shoot while the other does. By contrast consider a case like that in Figure 1 in which $U1= u11$ is realized by $L1=l11$ and we are interested in how $U1$ and $L1$ relate causally to $L2$ which we assume takes value $l22$.

The counterfactuals that correspond to (7.1-7.2) are:

7.3. If $U1= u11$ and $L1 \neq l11$, then $L2= l22$

7.4. If $U1 \neq u11$ and $L1 = l11$, then $L2= l22$.

The antecedent of (7.3) is possible (since $U1$ is multiply realizable and hence might have been realized by some other value of $L1$ besides $l11$). However (7.3) will not be true if there is a realizer of $U1$ (different from $l11$) which does not cause $L2=l22$. By contrast the antecedent of (7.4) is not possible. These differences between (7.1-2) and (7.3-4) reflect the fact that even if want to describe the case in which $L1$ realizes $U1$ as a case of overdetermination, it involves is a very different kind of overdetermination than is present in the riflemen case. Ordinary cases of overdetermination like the riflemen case are relatively rare and involve either "coincidences" or require the operation of some additional (ordinary) causal structure (e.g., an order from the commander to both riflemen to fire) the presence of which is contingent. This is why we don't think that such ordinary overdetermination is ubiquitous. By contrast the connection between $L1=l11$ and $U1= u11$ when the former realizes the latter is not a coincidence and not the result of some additional co-ordinating causal structure. This second sort of "overdetermination" is secured by the presence of the realization relation and for that reason it is both common and unmysterious. The argument that there is something puzzling or problematic about this second kind of overdetermination seems to rely on wrongly assimilating it to the first kind of overdetermination[29]

[28] These counterfactuals should of course be interpreted in an interventionist, non- backtracking manner.

[29] In thinking about overdetermination and "extra" arrows, it is also important to distinguish the question of which causal relations exist in nature from the question of which causal relations one *needs to represent* in a particular graph or other representational structure. Consider the usual case in which $L1$ causes $L2$ and $L1$, $L2$ realize $U1$ and $U2$, with $U1$ having a uniform effect on $U2$. If what we are interested in is explaining $U2$, we may legitimately decide to employ a graph in which there are no arrows from $U1$ to $L2$ or from $L1$ to $U2$ even if the effects are uniform. The reason for this is that the difference-making information in which we are interested is fully absorbed into the arrow from $U1$ to $U2$ – see Section 8. When we omit the arrows from

Finally, let me comment briefly on the argument that the postulation of downward (or indeed any upper-level causal relation) is superfluous or redundant, given the lower-level causal relationships. (This is in anticipation of some additional discussion in Section 8.) There is an obvious sense in which this claim of superfluousness is misleading, at least in the context of a situation like that described by Figure 1. The reason for this is that downward causation (or upper-level causation of upper-level effects) requires satisfaction of the conditions in Section **2** and these (and particularly the uniformity of effect requirement) are highly non-trivial. In particular, if *L1* causes *L2* and *U1* is realized by *L1* and *U2* by *L2,* it does *not* follow that *U1* has a homogeneous or uniform effect on *L2* (or on *U2*). Indeed if *L1* causes *L2*, then in the generic case, most ways of constructing upper level variable *U1* that involves coarse-graining *L1* will not yield variables that have a uniform causal effect on *L2* or on some *U2* constructed by coarse-graining *L2*[30]. In other words, given a diagram like Kim's, in which *L1* causes *L2*, it is wrong to think that if one countenances causation by upper-level variables at all, it follows automatically from the fact that *L1* realizes *U1* and *L2* realizes *U2* that one should draw additional arrows indicating causal relationships from *U1* to *L2* or from *U1* to *U2*. Again, one is entitled to do this only if the conditions described in Section **2** are met for these relationships. Thus when these conditions are met and on this basis we add arrows from *U1* to *L1* and/or from *U1* to *U2*, we are adding information to Figure 1 that does not follow just from the information in the lower half of the diagram — thus information that is not superfluous or redundant.

8. Conditional Causal Independence[31].

So far I have characterized a notion of downward causation, described some examples that I claim illustrate downward causation, and attempted to respond to several objections. However, an adequate defense of downward causation needs to do more than this; in particular, it would be desirable to have a more positive account of the *work* that is done by this notion— why it is a useful and fruitful notion in causal analysis, rather than, as critics claim, a dispensable and potentially confusing one[32]. In what follows I attempt to provide such an account, which appeals to a notion that I will call conditional causal independence. This will help us to better understand the worldly information that causal claims involving upper-level variables track.

U1 to *L2* or from *L1* to *U2* this need not be interpreted as claims that these causal relations do not exist; instead we have just declined to represent them.

[30] Note that even if *U1* has a uniform effect on upper - level variable *U2* it need not have a uniform effect on some lower-level variable *L2* that realizes *U2.* Suppose that the exact molecular state of a gas at time *t*, described by *L1*, causes its exact molecular state *L2* at some later time *t+d*, with *L1* realizing some upper level variable *U1*, e.g., temperature, at t. *U1* will not count as a cause of *L2* because it does not have a uniform effect on this variable, even though it may have a uniform effect on some upper level thermodynamic variable *U2* that is realized by *L2*.

[31] Here I want to acknowledge the influence of very similar ideas in Chalupka et al., 2017.

[32] Again, this follows from the idea that we want a "functional" account of causation—an account that shows how it is useful to think about causation in the way we do.

Suppose, as before, that we have an upper-level variable U the values of which are multiply realized by a lower level variable L (or set of these, but, as before, to simplify things I will assume that there is a single L) so that the L has a higher dimensionality than U . Let us say that L is *unconditionally causally relevant* to (alternatively, causally irrelevant to or independent of) some effect E if there are some (no) changes in the values of L when produced by interventions that are associated with changes in E. (Thus unconditional causal relevance is what is captured by the interventionist criterion for causation **M**). Say that L is causally irrelevant to (or independent of) *E conditional* on U if L is unconditionally causally relevant to E, U is unconditionally causally relevant to *E, and* conditional on the values of U, changes in the value of the L produced by additional interventions and consistent with these values for U irrelevant to E. In other words, we are to imagine a situation in which in which U and L are causally relevant to E, U is set to some value $u1$ via an intervention and then L is set via independent interventions to various values that are consistent with this value $U=u1$. If under such variations in L for fixed U, the value of E does not change, L is causally independent of E conditional on U[33]. For example, conditional on the setting of the temperature of a dilute gas to some value $T= t$, further variations in the kinetic energies of the individual molecules of the gas

[33] Some additional clarificatory remarks may be helpful. First, in contrast to the more familiar notion of conditional independence in probability theory, the notion of causal conditional independence is formulated in terms or interventionist counterfactuals—these rather than conditional probabilities provide the appropriate framework for understanding causal notions. Second note that we are *not* considering counterfactuals of the form: "If $L= l1$ and U were $= u2$, where $l1$ is not a realizer of $u2$, then…." As noted earlier such counterfactuals have impossible antecedents. Rather we are considering counterfactuals whose antecedents are, so to speak, the other way around, with the value $u1$ of U fixed and the L- realizers of that value $u1$ allowed to vary. These counterfactuals do have possible antecedents. Third, researchers who adopt the Stalnaker-Lewis closeness of possible world framework for evaluating counterfactuals sometimes argue as follows: Suppose that in the actual world, U takes the value $u1$, which is realized by $L=l1$, one of many possible realizers of U (the others being $l2$, $l3$…). Suppose we then consider a counterfactual whose antecedent is (1) "If L did not take the value $l1$, then…" It is then claimed that the possible world which is closest to the actual world in which the antecedent of (1) holds is one in which some other realizer of $U= u1$ obtains (that is, a world in which one of $L=l2$ or $L=l3$ etc. holds instead) and the counterfactual is evaluated accordingly. (Something like this idea is adopted in List and Menzies, 2009 to argue that true upper-level causal claims can *exclude* causal claims involving their lower-level realizers --- so called downward exclusion.) The framework described above does *not* rest on any such assumptions about closeness of worlds dictating which values of L would occur if $l1$ did not occur. Instead, we consider counterfactuals whose antecedents correspond to combinations of interventions where we specify exactly what is realized by those interventions rather than relying on closeness considerations to dictate what happens under those antecedents. (For a recent account of counterfactuals that exhibits these features see Briggs, 2012.) Thus when we consider counterfactuals like: if (i) we were to set $U=u1$ and independently of this (ii) set L to some other value (e.g. $l2$), different from $l1$ where $l1$ is the actual realizer of $u1$ but $l2$ is also a realizer of $u1$, we are not supposing that $l2$ would have been realized if $l1$ hadn't been. We are instead thinking in terms of a counterfactual the antecedent of which describes two separate operations, one of which sets $U=u1$ and the other of which sets $L=l2$.

as measured by some variable *K* when these variations are consistent with *T=t* will have the same effect (to a very high level of approximation) on other thermodynamic variables *E* such as pressure. Thus conditional on *T*, further variations in *K* are causally independent of such *E*s. Similarly, consider different lower-level ways *L= l1, l2, l3…* of realizing the same value of the membrane potential *V* in the HH model—these might correspond to slightly different distributions of charges in the membrane. Then we can interpret the HH model as claiming that conditional on the value of *V* (realized by an intervention), further variation in *L* whether *V* is realized by *l1* or *l2* or.. makes no difference to the channel conductances, so *L* is conditionally independent of these effects, conditional on the value of *V*.

When such a conditional independence relationship holds, *U* will of course have a uniform effect on *E*, regardless of how *U* is realized, and since by hypothesis *E* changes under some interventions on *U,* if *E* is a lower-level variable, *U* will meet the conditions in Section **2** for being a downward cause of *E*. One way of thinking about this is that under these conditions all of the information in the lower-level variable *L* that makes a difference for *E* is absorbed into the upper level variable *U* so that to the extent that explaining *E* is a matter of exhibiting those factors that make a difference for *E*, *U* does just a good a job in this respect as *L*. This justifies us in appealing to *U* as a cause of *E*. A similar analysis holds when *E* is an upper level variable.

This account has several additional features that are worth underscoring. Note first that we are assuming that *L* is unconditionally relevant to *E* as is *U*. Within the interventionist framework, this means that *both U* and *L* cause *E* so that, as remarked above (see footnote 32), we reject downward exclusion. The resulting "redundancy" is unproblematic, for reasons described in Section **7**.

A closely related point is that when a conditional independence relation of the sort described holds (with *L* being independent of *U* conditional on *E*) this by itself does not license the claim that upper-level causal claim provides a "better" explanation than the lower level claim. Rather what is licensed is the weaker claim that the upper-level explanation is just as good as the lower-level explanation as far as *E* is concerned—just as good because it captures all of the relevant difference-making information for *E* that is provided by *L*. This contrasts with the idea (accepted by many philosophers and some scientist who regard upper-level causal claims as legitimate) that the upper-level explanation in terms of *U* is superior to the lower-level explanation. This claim may be correct but it requires some additional argument for superiority.

Another point to keep in mind is that conditional causal independence relations are always relative to some target explanandum or effect *E*. That is, *L* might be conditionally causally independent of *E1*, given *U* but *L* might not be conditionally causally independent of some other explanandum *E2* given *U*. For example, there are many features of neuronal behavior which are dependent on the lower level details of exactly how charge is distributed along the neuronal membrane (again see Herz et al., 2006), even if this is not true for the effects described by the HH model.

9. The Role of Epistemic Factors

Considerations involving conditional independence of the sort just described can be invoked to explain why it is permissible or legitimate to formulate causal claims in terms of upper-level variables, including causal claims that involve lower-level variables as effects—when conditional causal independence holds we may lose little or nothing, in terms of difference-making information, by doing so. However, there is a crucial additional element to the story about why we actually employ such upper-level variables. This has to do with the various sorts of limitations that we humans (and perhaps all bounded agents) face. Some of these are calculational or computational -- we can't solve the 10^{23} body problem of calculating bottom up from the behavior of individual molecules to the aggregate behavior of the gas. Nor can we make the kinds of fine-grained measurements that would be required for such calculations to reach reliable results. Similarly, in the case of neuronal modeling, although there are more fine-grained models that describe the behavior of small individual "compartments" of the neuron , these cannot be simply "aggregated up" to produce a tractable model of the whole neuron (Again see Herz et al. 2006). We thus find that not only is it permissible to formulate theories in terms of upper level variables if we wish to explain certain explananda but that we have no alternative to doing so if we want models that are tractable or that we can calculate with. Put differently, we are very fortunate that nature presents us with relations of conditional irrelevance/independence of the sort I have been describing that we can exploit because otherwise scientific understanding of much or all of nature would be impossible. When we build models and theories that exploit these opportunities, we have models and theories in which upper level causation appears.

10. An Objection.

The ideas just defended are likely to prompt the following objection among reduction-minded critics. The objection is that on a view like mine top-down causation (and for that matter upper-level causation of upper-level effects) does not turn out to be "really real"—instead use of top-down causal claims just reflects shortcuts, approximations, idealizations etc. that scientists make for "pragmatic" reasons, like getting numbers out of their models, with genuine causation always occurring at a lower-level. Following this line of argument, it might be observed that in the HH model the neuron itself is composed of atoms and molecules which interact locally, mainly through the electromagnetic force. The membrane potential, the channel conductances and so on is thus the upshot or resultant of complex patterns of interaction among these atomic and molecular constituents. It follows, according to the argument we are considering, that *V*, the channel conductances and other variables in the HH model do not represent anything "over and above" these atomic constituents and their interactions. Similarly for the other putative examples of downward causation described above. We may be forced to talk in terms of causation by upper-level variables because of our computational and epistemic limitations, but (the objection goes) this just reflects something about us, not anything that is "out there" in the world or anything having to do with "what nature is really like".

There a number of things that might be said in response to this objection, many of which I lack the space to discuss. But one relevant consideration is this: the "world" and "what nature is like" do enter importantly into the account of downward causation that I have presented. That certain variables *L* are conditionally causally independent (or nearly so) of other variables *E*, given the values of other variables *U* is a fact about what the world is like, and not a fact about us or what we are able know or do. I see no reason to hold that facts about conditional independence are somehow unreal or in some way lacking in "objectivity". The way we should think about their status is not that our interests or limitations (or our willingness to employ pragmatic shortcuts) somehow create these facts about conditional causal independence. Rather the obtaining of these facts presents us with opportunities to formulate models and causal claims with certain structures (including those that contain claims of top-down causation) and which allow us to carry out calculations and construct derivations that would otherwise be impossible.

Another way of bringing out the role of these facts about conditional independence is to note that they are, so to speak effaced, if we focus only on derivations of particular token explananda from lower-level theory. To illustrate, consider a wildly counterfactual scenario[34] in which we are somehow able to deduce, from detailed information about the positions and momentum of each of the individual molecules making up a particular sample of gas and the fundamental laws governing their interactions—call this M1-- facts about the temperature and pressure of the gas, *E.* This deduction -- call it D-- by itself will not tell us which other microstates of the gas besides M1 would have led to *E* and which would have led instead to different values for the temperature and pressure. This last is information about conditional independence relationships and it is not apparent if we focus just on D. Of course, if we were somehow also able to derive for each possible set of values for the positions and momenta of the individual molecules, facts about the resulting temperature and pressure (i.e., if we could construct and survey all derivations of form D for all microstates of the gas), then this would tell us which microstates of the gas lead to *E* and which would lead to other values for the temperature and pressure. In this sense (it might be argued) information about the relevant conditional causal independence relations is "contained in" the representation provided by the lower-level theory, and, to repeat an earlier objection, not something that is "over and above" what is in this theory[35].

I accept this last claim, at least as far as the kind of realization relation on which I have focused in this essay is concerned. The defense of downward causation I have provided does not rest on claims about the emergence of novel causal facts that are somehow independent of all of the causal information (which I assume includes information about conditional causal independence) that follows in principle from the lower-level theory. Conditional independence of upper-level causal claims does not mean that those claims have no connection with lower-level

[34] Here I indulge a common claim in the philosophical literature: that all true upper-level claims are derivable in principle from information about lower-level variables and the laws governing their behavior. This claim should be treated with skepticism: One problem is that it is unclear, absent a specification of "in principle" and "derivable". If the derivation would require a computer as big as the solar system would that count as "in principle" derivability? And what counts as a "derivation"? For example, does it include use of limiting and asymptotic relations and perturbation techniques? Given that on many interpretations, quantum mechanics and quantum field theory give us only information about probabilities of outcomes, and that unlikely or unpredictable outcomes will sometimes occur and affect what happens later how does this affect such derivability claims?

[35] Even if it is true, as I am conceding for purposes of this essay, that we knew "everything" about the lower level variables and the laws that characterize their behavior and had unlimited computational power we could "derive" all true upper-level causal claims, it does not follow that the upper level claims are "reducible" to the lower-level claims. There are many accounts of reduction on offer but on most reduction requires something stronger than this sort of derivability. For example, many think it requires identities between upper-level and lower-level variables. As claimed previously, there are many cases identity is not the appropriate way to think about the relation between upper and lower. I will also add that "derive" in this claim requires a great deal of additional specification. For example, does it include limiting and asymptotic relations, perturbation techniques and so on?

theory. Again, I'm willing to stipulate, for purposes of this essay that these facts about conditional causal independence are in some way "contained in" the lower level theory. In this respect the account that I have provided respects what is sometimes called the causal closure of the physical—there is no invocation of causal facts that are not present in some form in the underlying physics.

However, leaving matters with just this observation leaves out some considerations of great importance, which have to do with epistemology of upper-level causal claims and methodologies for finding them The basic point is that finding out or "seeing" what information in lower-level causal claims is conditionally causally irrelevant to upper-level causal claims and what information is conditionally relevant and finding upper-level variables that capture conditional irrelevance relations is a highly non-trivial task. There are many other important issues concerning the relation between upper and lower-level causal claims besides the metaphysical ones reflected in denials or affirmations of "over and above" claims. The notion of conditional causal independence helps in thinking about these "other" issues

In an influential essay Anderson, 1972 notes there that even if it is true that all of the information that is relevant to some set of upper-level phenomena (such as superconductivity) is in some sense contained in an underlying theory, it may be as a practical matter difficult or impossible to extract the relevant variables for explaining these upper-level phenomena merely from an examination of the lower-level theory. One reason for this is that there are many different ways of forming upper-level variables from the variables of the lower-level theory and most of these will not lead to the successful formulation of conditional causal independence relations. The lower-level theory is not organized around (and doesn't care about) conditional causal independence facts involving upper-level variables, so that both upper-level information (e.g., empirically discovered regularities about superconductivity) and in some cases imaginative mathematical developments are required to find conditional causal dependence relations concerning upper-level variables. Reductivist minded philosophers sometimes neglect this because they think that the only relevant issue is whether various particular upper-level explananda are derivable in principle from lower-level facts. But as illustrated above, such derivations at least when considered individually, are not going to disclose the conditional independence relations and variables needed for the formulation of upper-level causal claims. And even putting this point aside, the fact of in-principle derivability tells us nothing about how to find the appropriate upper-level variables.

11. Autonomy.

The notion of "autonomy" is closely associated with issues having to do with the status of upper-level causal claims. What might it mean to claim that a set of upper-level causal claims are autonomous with respect to causal claims involving their lower-level realizers? One possibility is that upper-level causal claims are (in some sense) completely independent of the lower level causal facts—the upper-level is "novel" (and perhaps unexplainable even in principle) with respect to the lower level. I rejected this idea above. Another (in my view more reasonable) possibility is that autonomy has to do with the extent to which one can discover and formulate "good" upper level causal relationships without reference to information about their underlying realizers and the laws and causal relations governing these realizers. On this understanding of autonomy, the continuum mechanics of fluids is autonomous to the extent that one can formulate stable continuum level relationships with uniform effects of upper-level variables (e.g., as in the Navier-Stokes equations) without reference to underlying molecular details. Similarly, psychological generalizations are autonomous with respect to neurobiology to the extent that there are true psychological generalizations specifying uniform effects on other psychological variables, so that psychology can proceed independently of neurobiology. Of course the extent to which this sort of autonomy holds is an empirical matter.

This notion of autonomy is closely bound up with the extent to which various conditional independence relations hold, thus providing an additional illustration of the usefulness of the latter concept. When some causal claim featuring psychological variables is autonomous with respect neurobiology, then given the values of some psychological variables, further variation in the values of neurobiological variables will be causally irrelevant to other psychological variables, so that a conditional causal independence relation holds. When this is the case, we can ignore the neurobiology to the extent that we are interested in psychological effects[36]. Note again that this does not mean that the psychological claims are causally independent of the underlying neurobiology-- instead what is claimed is that the neurobiology is conditionally irrelevant to certain psychological variables, given other psychological variables[37]. Although I don't have the space to argue for this claim in detail here, I believe that it is only to the extent that such conditional causal independence relations hold that we have the possibility of upper-level or special sciences. This then is my answer to Fodor's well-known question, "why is there anything but physics?": The special sciences exist because or to the extent that the physics encodes conditional causal independence relations among variables that pertain to the sciences in question.

12. Is Conditional Causal Independence Common? Can We Make Sense of Closeness to Conditional Causal Independence?

Several philosophers with whom I have discussed this issue have claimed that causal conditional independence relations of the sort described (or even approximations to them), at least when they involve substantial reductions in degrees of freedom are very rare or perhaps non-existent and similarly for satisfaction for the conditions I have imposed on downward causation. Instead, their idea is that lower-level variables will always have a substantial causal impact on other variables, even conditional on the value of suitably chosen upper level variables. There are several things to be said about this. First, I emphasize again that whether conditional causal independence holds for various *L*s, *U*s and *E*s is always an empirical matter. It is plausible for some lower-level variables *L*, there may exist no *U*s with a substantial smaller dimensionality than the *L*s , conditional on which the *L*s become independent of explananda of interest. If the lower-level variables *L* in such a case have high dimensionality and/or the relations among them are highly complex, it may prove impossible to formulate true and non-trivial causal claims among upper-level variables. (Maybe some systems studied in the social sciences are like this.) Moreover, it *is* true, as noted above, that for most arbitrary sets of *U*s, *L*s and *E*s causal conditional independence, or even approximate causal conditional independence will fail. This

[36] Remember that this is a claim about what needs to be case for psychology to be autonomous from neurobiology and not an empirical claim about the extent to which such autonomy holds.

[37] It is also not claimed that if the causal relations among psychological variables are real, there must not be causal relationships among the underlying neurobiological variables which is what downward exclusion arguments claim.

does not in itself indicate anything about the usefulness of the conditional independence notion. It merely reflects that the fact that "good" upper level-variables are hard to find -- indeed, as noted above, they can be hard to find even given a lower-level ground truth from which the upper- level variables can be constructed.

Despite this I claim that in a number of cases, for *L*s that figure in an empirically well supported lower-level theory, there will exist upper-level variables *U* that render the *L*s conditionally causally independent of various explananda *E*s of interest. In other words, conditional causal independence or a close approximation to it sometimes holds, with the interesting scientific problem being to identify the variables for which it holds.

I have already mentioned some examples (Section 3) but here are some more general observations:

12.1. Physics. A number of physical systems exhibit universality in the sense of irrelevance of various sorts of lower-level detail to some aspects of the system's behavior, given certain upper level variables. For example, conditional on certain very generic or coarse-grained variables having to do with the symmetry of the system, its dimensionality, and the extent to which interactions are local, the lower-level details of many very different substances (different gas/liquid systems, ferromagnets etc.) are irrelevant to certain aspects of their behavior near their critical points. Explaining why this is so and identifying the relevant upper-level variables is one of the triumphs of the renormalization group analysis of such systems.

12.2. Biology. In many cases, organisms are constructed in such a way that certain variations in lower-level detail are conditionally irrelevant to more upper-level variables, given other upper-level variables to which the organism is responsive. This is so for a variety of reasons including selective pressures that reflect the desirability of eliminating the influence of various sorts of low-level noise, computational limitations which make it optimal for the organism to respond to coarse-grained variables and the fact that the coarse-grained variables can sometimes capture all that is ecologically significant. For example, it would make little sense for bodily responses of medium-sized organism like ourselves to dangerous stimuli to vary depending on the exact details of, say, the molecular realization of those stimuli—it is the fact that the stimulus is dangerous or perhaps that it involves a particular kind of danger (large predator) that is relevant. In such cases and for most of sensory processing we have screening off (conditional causal independence) of lower-level detail by ecologically relevant upper- level variables with respect to behavioral responses. Thus to the extent that organisms are only sensitive to coarse-grained variables rather the details of their realizers, good theories of the behavior of these organisms also may only have to keep track of coarse-grained variables. In general, there are many examples of biological systems in which some transducing system is sensitive only to lower dimensional patterns in some continuous lower-level variable with down-stream variables being influenced only by the information in the transduced pattern. Again in such cases one has conditional causal independence. That is, such systems operate by finding upper-level coarse grained variables that satisfy conditional independence relations with respect to lower level variables.

12.3. Relaxing Conditional Causal Independence. So far I have focused on cases in which complete conditional causal independence or something close to it holds. However, it is also worth exploring whether there are principled ways of relaxing that requirement[38]. One possibility is that although there may be rare or exceptional values of *L* that are conditionally relevant to *E*, even given the values of *U*, this may not be true for most or "almost all" values of *L* —for most or almost all such values, *L* is conditionally independent of *E,* given *U* even if there are a few values of *L* for which this is not true. Or perhaps conditional independence 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). Or conditional irrelevance or near conditional irrelevance may hold on some scales (typically coarser ones) but not on others. Yet another possibility is that when we consider possible probability distributions for the values of *L* that realize various values of *U* we find that conditional independence relations hold with respect to some *E* for most "well-behaved" probability distributions—e.g., those that satisfy some continuity condition.

Finally in cases in which we don't have complete conditional causal independence, a natural question to ask is how much explanatorily or causally relevant information about *E* do we "lose" if we employ *U* instead of *L*? (Here relevant information is information about difference-making variables, understood along interventionist lines.) One possible way of doing this employs a notion of conditional mutual information interpreted causally along the lines described in Ay and Polani, 2008: the information loss if we employ *U* instead of *L* (or gain if we employ *L* instead of *U*) is measured by $I\ (E{:}\ L)\ |U)$ the mutual information between *U* and *E* conditional on *L* where *U* and *L* are set by independent interventions in the manner described above[39]. Complete conditional causal independence then corresponds to the case in which $I\ (E{:}\ L)\ |U)= 0.$

References

Anderson, P.W. (1972) "More is Different" *Science* 177:393-396.

Ay, N. and Polani, D. (2008) "Information Flows in Causal Networks" *Advances in Complex Systems* 11, 17-41.

Baumgartner, M. (2010), "Interventionism and Epiphenomenalism" *Canadian Journal of Philosophy* 40, 359-383.

Baumgartner. M (2018). "The Inherent Empirical Underdetermination of Mental Causation". *Australasian Journal of Philosophy.* 96: 335-350.

Bechtel, W and Craver, C (2007) "Top-down Causation Without Top-Down Causes" *Biology and Philosophy* 22:547–563.

Beckers S. and Halpern, J. (2019) "Abstracting Causal Models" **arXiv:1812.03789 [cs.AI]**

[38] Of course this needs to be done in such a way that "highly" ambiguous interventions are avoided.

[39] Ay and Polani 2008 employ this expression to measure what they call "information flow" between variables in an ordinary causal network with no realization relations present. I suggest that the same measure can be used to measure conditional information when realization relations are present.

Blanchard, T., Murray, D. and Lombrozo, T.
(Forthcoming). "Experiments on Causal Exclusion". *Mind and Language.*

Briggs, R. (2012) "Interventionist Counterfactuals" *Philosophical Studies* 160: 139-166.

Butterfield, J. (2012) "Laws, Causation and Dynamics at Different Levels" *Interface Focus* 2, 101–114.

Chalupka, K. Eberhardt, F., and Perona, P. (2017) "Causal Feature Learning: An Overview" *Behaviormetrika* 44:137–164.

Clark, S. and Lancaster, T. (2017) "The Use of Downward Causation in Condensed Matter Physics". In Paoletti, M. and Orilia, F (eds.) *Philosophical and Scientific Perspectives on Downward Causation*. New York: Routledge, 42-53

Ellis, G. (2016) *How Can Physics Underlie the Mind? Top-Down Causation in the Human Context*. Berlin: Springer.

Herz, A. Gollisch, T . Machens, C. Jaeger, D. (2006) "Modeling Single-Neuron Dynamics and Computation: a Balance of Detail and Abstraction" *Science* 314, 80- 85.

Kim. J. (2005) *Physicalism or Something Near Enough.* Princeton: Princeton University Press.

Lewis, D. (1986) *Philosophical Papers, Volume II* Oxford: Oxford Univeristy Press.

Lewis D. (2000). "Causation as influence". Reprinted in Collins J., Hall N. and Paul L. (eds), *Causation and Counterfactuals*. Cambridge: MIT Press.

List, C. and Menzies, P. (2009). "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106: 475-502.

Noble, D. (2006) *The Music of Life.* Oxford: Oxford University Press.

Pearl, J. (2000) *Causality.* Cambridge: Cambridge University Press.

Rubenstein, P. , Weichwald, S.; Bongers, S.; Mooij, J. Janzing, D.; Grosse-Wentrup, M.; and Scholkopf, B. (2017). "Causal Consistency of Structural Equation Models". In *Proc. 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*.

Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction and Search*. Cambridge: MIT Press.

Spirtes, P. and Scheines, R. (2004) "Causal Inference of Ambiguous Manipulations" *Philosophy of Science* 71:833-845.

Woodward, J. (2003) *Making Things Happen*. New York: Oxford University Press.

Woodward, J. (2008) "Mental Causation and Neural Mechanisms" In Howhy. J. and Kallestrup, J. , *Being Reduced: New Essays on Reduction, Explanation and Causation.* Oxford: Oxford University Press, 218-262.

Woodward, J. (2014) – "A Functional Account of Causation" *Philosophy of Science* 81: 691-713.

Woodward, J. (2015) "Interventionism and Causal Exclusion" *Philosophy and Phenomenological Research* 91: 303- 347.

Woodward, J. (2017) "Interventionism and the Missing Metaphysics" In *Metaphysics and the Philosophy of Science* (ed. Mathew Slater and Zanja Yudell), pp 193-227.

Woodward, J. (2020) "Levels: What are They and What Work Do They Do?" in Kendler and Parnas ed. *Philosophical Issues in Psychiatry V: The Problems of Multiple Levels, Explanatory Pluralism, Reduction and Emergence.* Cambridge: Cambridge University Press.

Woodward, J. (Forthcoming) *Causation with a Human Face: Normative Theory and Descriptive Psychology.* New York: Oxford University Press.