Synthese (2021) 198:237–265
https://doi.org/10.1007/s11229-018-01998-6

Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance

James Woodward[1]

Received: 19 June 2017 / Accepted: 20 October 2018 / Published online: 31 October 2018

Abstract
This paper responds to recent criticisms of the idea that true causal claims, satisfying a minimal "interventionist" criterion for causation, can differ in the extent to which they satisfy other conditions—called stability and proportionality—that are relevant to their use in explanatory theorizing. It reformulates the notion of proportionality so as to avoid problems with previous formulations. It also introduces the notion of conditional independence or irrelevance, which I claim is central to understanding the respects and the extent to which upper level explanations can be "autonomous".

Keywords Interventionism · Proportionality · Stability · Conditional irrelevance

1 Introduction

Woodward (e.g., 2008, 2010) defended the view that true causal claims, satisfying a minimal "interventionist" criterion for causation (Sect. 3 below), can differ in the extent to which they satisfy other conditions relevant to their use in explanatory theorizing—these include *stability* and *proportionality,* among others. The stability of a causal relationship has to do with whether it would continue to hold under changes in background conditions. Proportionality, on the interpretation described below, has to do with the extent to which a causal claim fully captures conditions under which variations in some phenomenon of interest occur. Depending on empirical considerations and our target explananda, causal claims at different "levels"[1] or framed in different vocabularies may satisfy stability and proportionality requirements to differing degrees and thus differ in their appropriateness or informativeness for explanatory purposes. This

[1] The use of words like "level", "upper", and "lower" is ubiquitous in the philosophical literature, including Weslake's and Franklin-Hall's papers. I adopt this usage for convenience, even though it is problematic. I favor a very deflationary reading: talk of levels is just a way of expressing (local) claims about conditional independence, in the sense of this notion described in Sect. 5.

✉ James Woodward
jfw@pitt.edu

[1] Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, USA

can help in finding the most appropriate level (and associated vocabulary) at which to frame explanatory claims. Sometimes this framework favors "upper-level" over lower-level explanations, although in other circumstances, depending on the empirical facts, it favors lower-level explanations.

Recently these ideas have been criticized by several writers. Franklin-Hall (2016) criticizes Woodward's formulation of proportionality (as well as, by implication, other formulations in the literature such as Yablo (1992)). She also argues that, even when suitably reformulated, proportionality cannot be used to motivate a "non-pragmatic" preference for upper-level explanations over those provided by some lower-level theory. (See Sect. 2 for what "non-pragmatic" means.) Related objections to proportionality have been advanced by Shapiro and Sober (2012), and Maslen (2009). Both Franklin-Hall and Weslake (2010) also object to Woodward's appeal to stability considerations in support of upper-level explanations. They argue that a focus on stability, combined with other features of Woodward's account, instead leads to the conclusion that upper-level explanations are always "non-pragmatically" *inferior* to those provided by lower-level theories—particularly those of "fundamental physics".[2]

This paper responds to these objections and also attempts to place them within a more general framework for thinking about the status of upper-level explanations. I begin (Sect. 2) with some brief remarks about the overall strategy of Weslake's and Franklin-Hall's arguments. I then turn to the notion of proportionality, arguing, in agreement with Franklin-Hall, that current formulations of this notion including the formulation in Woodward (2008, 2010), are inadequate. Sections 3, 4 reformulate this notion and make its underlying rationale more transparent. Section 5 then introduces the notion of conditional irrelevance, which I argue is central to the justification of upper-level explanations and to whatever "autonomy" they possess. Section 6 discusses an alternative proposal by Weslake concerning what the "non-pragmatic superiority" of upper-level explanations consists in.

2 Background: non-pragmatic superiority and W-questions

I begin with some remarks about the overall strategy of Weslake's and Franklin-Hall's arguments. This will be crucial to some of the more detailed arguments that follow. As noted, both authors organize their discussions around the question of whether it is possible for "upper-level" explanations to be "non-pragmatically" superior to lower-level explanations of a sort that (they suppose) are provided by fundamental physics. They agree, as does virtually everyone, that there are "pragmatic" reasons (where these are taken to include reasons having to do with our epistemic and calculational limitations—hereafter just "limitations") that support employment of "upper-level" explanations over lower-level alternatives. Because of these limitations, we cannot actually construct or exhibit explanations of many features of upper-level phenomena (e.g., the behavior of the stock market or short term memory) from fundamental physics. They ask, however, whether in addition to these pragmatic advantages, there

[2] For additional discussion of Franklin-Hall on proportionality and a defense of interventionism, see Blanchard (forthcoming). I see Blanchard's discussion as complementary to mine.

are further non-pragmatic features that show upper-level explanations to be superior. In effect, they consider situations in which we have available a complete fundamental physics and in which we abstract away from all epistemic and calculational limitations standing in the way of constructing explanations of upper-level phenomena from fundamental physics. They then ask whether in this situation, there would be *additional* reasons of a non-pragmatic sort for preferring upper-level explanations. Weslake argues that there are such additional reasons, calling this view *Autonomy* and contrasting it with its denial, *Fundamentalism*. Weslake claims that interventionism is committed to *Fundamentalism* and criticizes it accordingly. Franklin-Hall does not explicitly argue that there are non-pragmatic reasons why upper-level explanations are superior to fundamental explanations but she describes this as a widely shared view and faults interventionism for failing to capture it.

An important part of the background to this discussion is Woodward's (2003) and Hitchcock's and Woodward's (2003) accounts which connect the "explanatory depth" of an explanation to its ability to answer a range of questions (w-questions) about the circumstances in which its explananda would have been different. In doing this, the explanation exhibits a pattern of dependence between explanans and explanandum.[3] My view, as argued below, is that this idea can be thought of as helping to provide an underlying motivation for the use of proportionality and stability in explanatory assessment and hence in finding an appropriate level for framing upper-level explanations. By contrast, both Weslake and Franklin-Hall argue that, putting aside pragmatic considerations, explanations in terms of fundamental physics always answer more w-questions than upper-level alternatives and hence (mistakenly) will be judged superior by the w-question criterion.

I find this overall line of argument unpersuasive for a number of reasons. Here I focus on just one of these, which is crucial to understanding how the w-question criterion (as well as the notions of proportionality and stability) should be interpreted. When interventionism speaks of an explanation *answering* a wide range of w-questions it means just that—that the explanation actually *displays* or *exhibits* a pattern of dependence that provides answers to such questions, showing us how changes in the factors cited in the explanans are associated with changes in the explanandum. I will not try to provide a complete account of what "*displays* or *exhibits*" means since what will matter most for my discussion is that we be able to recognize examples in which there is a clear failure of this condition. However, the basic idea is that the candidate explanation should provide information from which one can read off or access the claimed pattern of dependency of the explanandum on the explanans. In physics this typically involves exhibiting a derivation or calculation or solving an equation, either analytically or via perturbation methods, where this connects the explanandum with the explanatory premises one employs. On the interventionist view, there is thus an important difference between, on the one hand, the claim (supposing this to be true) that such a derivation or an answer to a w-question "exists in principle" or is "implicit" (Weslake 2010) in some theory and, on the other hand, actually writing down an explicit set of governing equations and exhibiting solutions to them or displaying

[3] Here the relevant notion of explanation is explanation of *why* some explanandum obtains, as opposed to explanations that answer who- or what-questions.

the steps in a derivation leading from the explanans to explanandum which show how the latter depends on the former. On the interventionist view, only the latter counts as "exhibiting" or "displaying" an explanation or "providing" answers to w-questions. For example even assuming that the claim that the behavior of the stock market is "implicit" in the standard model of fundamental physics or is "in principle derivable" from this model is true, such a claim does not, in the relevant sense, provide answers to w-questions about stock market behavior and does not amount to the exhibition of an explanation for such behavior.[4]

In contrast to the situation envisioned above, in other contexts the exhibition of a pattern of dependency may simply involve the presentation of a causal claim. To use an example discussed below, the causal claim

(2.1) The presentation of a red target causes a pigeon to peck

will be naturally interpreted by many as implying that a claim that a pattern of dependence exists between whether or not the target is red and whether or not the pigeon pecks. I take this to be a case in which (2.1) exhibits a pattern of dependence because this pattern is readily accessible to those who understand (2.1). This claim of accessibility is supported by recent psychological research—see, e.g., (Sloman and Lagnado 2005) which shows that people readily associate claims like (2.1) with interventionist counterfactuals expressing dependency relationships.

Returning to the distinction between establishing that (1) an explanation exists, in the sense in which it might be claimed, e.g., that an explanation of stock market behavior in terms fundamental physics "exists" or is "implicit" in fundamental physics and (2) actually displaying or exhibiting an explanation of stock behavior, I take it to be uncontroversial that there is some distinction of this sort that can be drawn. However, it is a further question why it *matters* whether a researcher accomplishes (2) rather than just establishing (1). Why adopt an account of explanation that is sensitive to this distinction? One important reason is that (2) is an important goal of inquiry in its own right that is not achieved just by establishing (1). Even if it is true that, abstracting from our limitations, answers to questions about stock behavior are derivable in principle from fundamental physics, no serious researcher would think that just observing or

[4] Note that the distinction between, on the one hand, (1) merely claiming that the possibility of an in-principle derivation is implicit in some theory and, on the other, (2) writing down an explicit model, exhibiting solutions to the equations that figure in it, and/or exhibiting a derivation of the explanandum is an "objective" difference that does not depend on people's interests, abilities or opinions. Our interests or goals lead us to *care* about (2) in addition to (1) but that does not mean that the difference between (1) and (2) is subjective or interest-dependent. Note also that taken in itself, the distinction between (1) and (2) does *not* coincide with the distinction between derivations or calculations that humans are able to follow and those that they are not able to follow. (2) can be satisfied even if humans are unable to follow the exhibited derivation. Moreover, even if those explanations we in fact produce or exhibit *are* influenced by what we are able to calculate or keep track of, it still does not follow that there is no difference between (1) and (2) or that this difference is in some way a "subjective" or "anthropomorphic" matter. Put differently if, say, we regard considerations having to do with what we are able to calculate as "pragmatic" and allow that these influence the explanations we construct and exhibit, it again does not follow that the distinction between (1) and (2) is "merely pragmatic"— at least if "pragmatic" is interpreted to mean "subjective" or "arbitrary". In the same way, there is an objective difference between claiming, however truly, that a proof for some mathematical claim exists and, alternatively, exhibiting or producing such a proof, and this is so, even if the form taken by the proof is influenced by what we are able to comprehend or follow—this does not make it a non-objective (or even in any clear sense a "pragmatic") matter whether a valid proof has been exhibited.

establishing that this derivability claim is true is an end point in inquiry. Instead a central goal of researchers is to *exhibit* factors on which the movement of stock prices depends and to explicitly connect the movement of stock prices to these factors. Our current science is such that doing this requires "upper-level" theories of the sort found in economics and finance, rather than appeals to fundamental physics. Some philosophers may think that this is a temporary state of affairs but even if this turns out to be true (a possibility I would regard as extremely unlikely), it remains true that these upper level theories do or at least aspire to do something [described by (2)] that is both valuable and distinct from merely establishing that explanations "exist".

With this motivation in mind, I will follow Woodward (2003), (especially pp. 157–161, 175–181, see also Woodward 2016b) in understanding the notion of explanation in such a way that an explanation is something that displays or exhibits dependency relations.[5] This follows ordinary usage (no researcher thinks it is an explanation of the behavior of the stock market to observe that such behavior must be derivable in principle from fundamental physics) but more importantly reflects the idea, developed in more detail in Sect. 5 and "Appendix 1", that we should think of explanation as a goal that can play a methodological role in guiding inquiry and in theory construction.[6] It is exhibited explanations that can function in this guiding role.

As another illustration of this idea, consider Richard Feynman's remark in the mid-1950s that he did not understand why superconductivity occurred (Anderson 2011). Feynman made this remark despite having helped to create the fundamental theory (QED) which governs the behavior of electrons and which (he knew) describes the fundamental physics that "underlies" superconductivity and from which, given the assumptions that Weslake and Franklin-Hall make, superconductivity is presumably "derivable in principle". As I understand the notion, an "explanation" of superconductivity requires much more than the truth of this derivability in principle claim—it requires actual exhibition of difference-making relations relevant to the upper-level behavior (as in, for example, the Bardeen, Cooper, and Schrieffer theory). It is this that Feynman did not possess and that Bardeen, Cooper and Schrieffer won the Nobel Prize for.

More generally, in what follows I will understand the notion of an explanation answering a w-question in such a way that only explanations that exhibit answers to such questions count as providing answers to them. Similarly stability and proportionality will be understood as criteria for choosing among explanations and causal representations that we are able to exhibit or formulate, rather than criteria for the evaluation of supposed explanations that will never be exhibited.

[5] Readers who do not like this proposal can simply keep in mind the distinction between a displayed explanation and the claim that an explanation exists and take my subsequent discussion to be concerned with the former.

[6] Despite these remarks, some readers have suggested I have in some way confused establishing or claiming that an explanation "exists" with our ability to produce or display this explanation. I hope that my distinction between (1) and (2) above makes it clear that I have not fallen victim to any such confusion. My guess is what is really bothering such readers is not that I fail to distinguish (1) and (2) but rather their suspicion that (2) does not matter over and above (1) or that to the extent that it does matter, this has to do with "mere pragmatics", so that philosophical discussion should focus just on (1) and not concern itself with (2). This may well be Franklin-Hall's and Weslake's view. I would reject it for reasons described in the text.

There is more to be said about non-pragmatic superiority and about how the w-condition criterion should be understood but in order not to obstruct the flow of my discussion, I have relegated additional discussion to two appendices.

3 Interventionism and proportionality

Consider a causal claim of form "X causes Y in background circumstances B" where "causes" is understood to mean that a type level relation of causal relevance holds between X and Y, where X and Y are variables. The following is a slight modification of a proposal in Woodward (2008) which I label (**M***)[7]

> (**M***) X causes Y in B if and only if there are distinct values of X, x_1 and x_2, with $x_1 \neq x_2$ and distinct values of Y, y_1 and y_2 with $y_1 \neq y_2$ and some intervention such that if that intervention were to change the value of X from x_1 *to* x_2, then Y would change from y_1 to y_2.[8]

[7] (**M***) represents one of several possible choices in an interventionist treatment of causation. An alternative, stronger interventionist condition is this:

> (**M****) X causes Y in B if and only if there are distinct values of X, x_1 and x_2, with $x_1 \neq x_2$ and distinct values of Y, y_1 and y_2 with $y_1 \neq y_2$ such that under all interventions in B which change the value of X from x_1 *to* x_2, Y would change from y_1 to y_2.

The difference between (**M***) and (**M****) is that (**M****) replaces the reference to *some* interventions in (**M***) with a reference to *all* interventions. (**M****) requires that there be values of X, x_1 and x_2 such that under *all* interventions that change X from x_1 to x_2, Y changes uniformly from y_1 to y_2. Note, however, that (**M****) requires that this be true only for *some* pairs of values of X and Y, not for all such pairs of values. This last observation becomes important when we consider variables that are not binary. Suppose X has three possible values, x_1, x_2 and x_3 and Y three possible values y_1, y_2 and y_3. Then (**M****) will be satisfied as long as, e.g., all interventions that change x_1 to x_2 change Y from y_1 to y_2 even if interventions that change X to x_3 do not change the value of Y or sometimes change it and sometimes do not.

Suppose that we take the variable in the cause position of (3.1) below to take the values {scarlet, non-scarlet}. Then although (**M***) counts (3.1) as true, (**M****) counts (3.1) as false. Although it is true that, given the causal structure of the pigeon's situation, all interventions that set the target color to scarlet is followed by pecking, it is not true that all interventions in B that set the target color to non-scarlet are followed by non-pecking, since some of these interventions will involve setting the target color to some non-scarlet shade of red which will be followed by pecking. If (3.1) is false, there is of course no puzzle about why (3.2) is preferable to (3.1)—we don't need to appeal to proportionality to explain this. However, there are a number of other examples, described below, that strongly suggest that satisfaction of a plausible proportionality requirement should not be regarded as a necessary condition for a causal claim to be true even if one holds that (**M****) is the right account of the truth conditions for causal claims—that is, a causal claim can satisfy (**M****) (as well as (**M***)) and hence be true, even though the claim can fail to satisfy or fully satisfy a plausible version of proportionality. In part for this reason, it will make little difference to the overall structure of my argument whether (**M***) or (**M****) is adopted and in what follows I will generally adopt (**M***).

My own view is that there is no clear sense in which either (**M***) or (**M****) is more "correct"—I see them as alternative ways of regimenting causal language, each with advantages and disadvantages. In favor of (**M****) it might be argued that if, e.g., some interventions that set the target to non-scarlet lead to pecking and others to non-pecking, this shows that the intervention is ambiguous in the sense of Spirtes and Scheines (2004) and hence that the associated counterfactual is false. In favor of (**M***) is the fact that (**M****) is very demanding and appears to count as false many causal claims that we ordinarily think of as true, such as the Sober/Shapiro example "$X=3$ caused $Y=6$" discussed below. Thanks to [reference omitted] for very helpful discussion.

[8] Purely for reasons of expository convenience, I will assume that the systems with which we are dealing in this paper are deterministic, so that there is always a determinate answer to the question of how if at all

"Intervention" should be understood along the lines described in Woodward (2003).

When a variable X satisfies (**M***) with respect to Y, I will sometimes say that X is a "difference-maker" for Y and that there is a "dependency relation" between X and Y. I will also use the phrases "causal claim" and "causal explanation" interchangeably in what follows, assuming that a causal explanation for some variable Y is simply an assembly of information about the causes of Y (understood in accordance with (**M***)) perhaps along with a more or less explicit statement of how Y depends on these causes, along the lines described in Woodward (2003) and Hitchcock and Woodward (2003).[9]

As Woodward (2008) notes, (**M***) provides, at best, a kind of minimal condition for causation, in the sense that a claim can satisfy **(M*)** and yet seem, in other respects, less preferable (less perspicuous, informative, illuminating or explanatory) than other true causal claims, also satisfying (**M***), that might be formulated about the system of interest.[10] In other words, different causal claims can each satisfy **(M*)** and yet differ along other, additional dimensions apparently relevant to their assessment. *Proportionality* is one such dimension.

Turning now to this notion, Yablo's informal gloss is that proportionality requires that causes be "just enough" for their effects, neither omitting relevant detail nor being overly specific in the sense of containing irrelevant detail.[11] (We will see shortly that this motivating idea is potentially misleading.) In one of Yablo's illustrations,[12] a pigeon is trained to peck at targets of any shade of red and only such targets. The pigeon is presented with a scarlet target and pecks. Consider the following two claims:

(3.1) The scarlet color of the target causes the pigeon to peck.
(3.2) The red color of the target causes the pigeon to peck.

Footnote 8 continued
Y would change under an intervention on X. However, (**M***) may be readily extended to stochastic systems by talking about whether a change in the probability distribution of Y would occur under an intervention X.

[9] Parallel remarks apply to (**M****). (See footnote 7.)

[10] Note that a parallel remark applies to (**M****)**:** (**M****) is satisfied as along as there is some pair of values x_1, x_2, and y_1, y_2 $y_1 \neq y_2$ such that all interventions that change x_1 *to* x_2, change Y to y_1 to y_2. A causal claim can satisfy this condition and be uninformative about what would happen to Y under changes in other values of X. In such a case the causal claim will still be less informative than ideally we would like it to be and we need a notion like proportionality to capture this. This is one reason why, as suggested earlier, it makes little difference to my overall argument of we adopt (**M****) rather than (**M***).

[11] Yablo's more precise characterization is this: The having of property C is proportional to effect E if and only if (1) for any determinable C* of C, had C* obtained without C, E would not have obtained, and (2) for any determinate C′ of C, had C obtained without C', E would still have obtained. Yablo (1997, pp. 267–268) also formulates this idea in terms of "screening off" relationships between determinables and determinates, presumably in analogy with the screening-off relations employed in discussions of probabilistic causation, although the latter have to do with probabilistic independence, while Yablo makes use of a notion of conditional counterfactual independence, as I do below (Sect. 5). The characterization I provide, though, differs from Yablo's by being framed in terms of variables, which may be non-binary as well as binary, rather than properties. I also depart from Yablo in thinking of proportionality as a matter of degree. Nonetheless, I think, as will become apparent below, that the general idea behind this formulation (which I interpret as a kind of conditional counterfactual (in)dependence condition) captures a very important feature of causal and explanatory thinking and that Yablo's introduction and development of this idea is an important achievement.

[12] I use this example because it has been widely used in discussions of proportionality. For illustrations of the use of proportionality that are more scientifically serious, see Woodward (2010).

Yablo claims (3.2) is superior to (3.1) on the grounds that (3.1) is "overly specific" and fails to be "proportional" to its effect. Yablo does *not* take proportionality to be a necessary condition for the truth of causal claims (1992, p. 277) but a number of other writers either interpret Yablo as claiming this (e.g., Shapiro and Sober 2012) or advocate this position themselves (e.g., List and Menzies 2009).

To represent (3.1) and (3.2) within an interventionist framework we need to express them in terms of claims about variables. (3.2) does not explicitly display the alternative possible values of the variables figuring in it but it is very natural to interpret it as employing a cause variable *RED* that is conceptualized as a binary variable that can take either of two values, {red, not-red}, and an effect variable *PECK* conceptualized as a binary variable capable of taking the values {peck, not-peck}. (3.2) might then be represented as:

(3.3) *RED* causes *PECK*

where this is unpacked as implying that

(3.3a) An intervention that sets *RED*=*red* is followed by *PECK*=*peck* (alternatively, *RED*=*red* causes *PECK*=*peck*)

(3.3b) An intervention that sets *RED*=*not red* is followed by *PECK*=*not peck* (alternatively, *RED*=*not red* causes *PECK*=*not peck)*

Obviously (**M***) counts (3.2)–(3.3) as true.

How about (3.1?) One possible way of representing it within an interventionist framework is to employ, along with *PECKS*, a variable *SCARLET* that takes the values {scarlet, not scarlet}. On this interpretation, (**M***) also counts (3.1) as true. This is because, given the causal structure of the pigeon's situation, an intervention that sets the target color to scarlet is followed by pecking and there exists an intervention that sets the target to some different color (e.g., blue) that is followed by non-pecking. In my idiolect, I'm inclined to count (3.1) as true, in agreement with the judgment of (**M***). Other examples, described below, suggest even more strongly that satisfaction of a proportionality requirement should not be regarded as a necessary condition for a causal claim to be true. Nonetheless it seems to me that Yablo is clearly correct in thinking that (3.1) is in some way inferior to or less perspicuous than (3.2) in describing the pigeon's situation. I take this to motivate the general idea that (3.1) is deficient but that this deficiency is not a matter of (3.1) being false. Instead, following Woodward (2010), I will take this deficiency to have to do with an explanatory limitation in (3.1)—it is inferior qua causal *explanation* in comparison with (3.2) and fails to convey or exhibit explanatorily relevant information in comparison with (3.2).

To further spell out this idea, note that, intuitively, within an interventionist framework, there are at least two different ways in which a causal explanatory claim might be deficient:

(1) It might falsely claim that some dependency relationship is present when it is not—call this *falsity*.
(2) It might fail to represent dependency relations that are present—call this *omission*.

When (3.1) is interpreted (in accord with (**M***)) as the claim that there exist interventions in which the target is set to non-scarlet which are followed by non-pecking it does

not make false claims about the existence of dependency relations. However, under this interpretation, there are facts about how the occurrence of pecking depends on the color of the target (in particular that pecking occurs for non-scarlet but red targets) that are not conveyed by (3.1). In this sense there are dependency relations present in the example that (3.1) fails to capture or represent, so that (3.1) when so interpreted exhibits failures along the omission dimension (2). I will suggest below that many failures of causal claims to fully satisfy proportionality have this feature—they involve claims that correctly represent some dependency relations that are present but that fail to represent others.

Since the extent to which an explanation fails to convey information about dependency relations that are present seems in many cases to be a matter of degree, we should expect this also will be true of proportionality—it too will come in degrees, rather than being an all or nothing matter. (Since truth and falsity don't come in degrees this is another reason why proportionality should not be regarded as a necessary condition for the truth of causal claims.)

In an attempt to capture this line of thought, Woodward (2010) proposed the following proportionality-like condition (meant to capture the extent to which a causal or explanatory claim satisfies proportionality). Unlike Yablo's formulation, this was intended to apply to causal claims that relate non-binary variables and that are not necessarily related as determinates and determinables. Causal claims also can vary in the degree to which they satisfy (**P**):

> (**P**) There is a pattern of systematic counterfactual dependence (with the dependence understood along interventionist lines) between different possible states (or values) of the cause and the different possible states (values) of the effect, where this pattern of dependence at least approximates to the following ideal: the dependence (and the associated characterization of the cause) should be such that (a) it explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys only such information—that is, the cause is not characterized in such a way that alternative states of it fail to be associated with changes in the effect. (Woodward 2010: p. 298)

One way in which (**P**) is inadequate is brought out by the following counterexample due to Franklin-Hall (2016). Suppose the causal facts involving the pigeon are as described earlier and consider a new binary variable V, which has two values, scarlet and cyan, with the pigeon pecking when the target is scarlet but not when it is cyan. Consider

(3.4) V's taking the value = *scarlet* causes *PECK*=*pecking*

(3.4) satisfies both (a) and (b) in **P.** But (3.4) seems defective in comparison with (3.2) and also seems to violate what Franklin-Hall calls the "spirit" of proportionality, which, following the discussion above, I take to involve the extent to which dependency relations that are present are represented. (3.4) fails to represent some dependency relationships that are present (those involving red but non-scarlet and non-cyan but non-red targets) despite satisfying (**P**). Note also that (3.4) also satisfies both (**M***)

as well as (**M****) (footnote 7)—thus further illustrating the need for an additional condition on explanatory assessment that goes beyond (**M***) and (**M**).**

Another counterexample is due to Shapiro and Sober (2012). Their immediate target is the following characterization of proportionality:

> A statement of the form 'C caused E' obeys the constraint of proportionality precisely when C says no more than what is necessary to bring about E. (p. 89)

They consider a case in which real-valued variables X and Y are related by some non-monotonic function F which maps two different values of X—e.g., 3 and 22—into the same value of Y ($y=6$), with other values of X being mapped into different values of Y. There is an obvious sense in which $X=3$ is not "necessary" (or is "overly specific") for $Y=6$. So proportionality, understood as above, fails in this case. Assuming (as Shapiro and Sober do) that proportionality is put forward as necessary for causation, it seems absurd to deny that, in the example as described, $X=3$ caused $Y=6$ or to claim that the relationship between X and Y is not a causal relationship. (Note that this relationship is causal according to (**M***).) Moreover, Sober and Shapiro's example seems to tell equally against clause (b) in (**P**): it seems misguided to contend that a causal claim described by a non 1–1 function is for that reason less preferable or explanatory than a causal claim described by some 1–1 function, but the requirement that "alternative states of the cause are associated with changes in the effect" seems to have just this implication.

Note that it is natural to think about (3.1)–(3.2) as well as (3.4) as raising issues about *variable choice*—about which variables we should employ in formulating causal relationships. (For example, V in (3.4) is, intuitively, a less "good" variable than *RED*.) As we see from these examples, different variables can lead to the formulation of different causal claims that, even if true, differ along the *omission* dimension (2) described above. I take one of the intuitions behind proportionality to be that, other things being equal, we should choose variables that allow for the formulation of causal claims that, in addition to satisfying (**M***), do a better rather than a worse job of satisfying (2).

Now consider another example. Suppose that smoking S causes both lung cancer L and yellow fingers Y but that neither L nor Y cause the other. Suppose we represent this causal structure as:

$S \rightarrow Y$

Unlike (3.1) and (3.4) the problem is now not that we have the "wrong" variables but rather that we have omitted a variable and an arrow ($S \rightarrow L$). In this way we failed to represent a dependency relation that is present. Alternatively, suppose that we adopt a representation in which there is an arrow from Y to L. Here the mistake is a representation of a non-existent dependency relationship, but again this is not because we have the wrong variables.

Discussions of proportionality in the philosophical literature have generally focused on examples like (3.1) versus (3.2 or 3.4), in which problems arise because of deficient variable choice (and in particular, choice of variables with the wrong grain, leading to failure to accurately represent dependency relations). In the interest of consistency I will follow this, distinguishing failures of proportionality (which I will take to always involve failures associated with variable choice) from omitted arrows. I acknowledge

some arbitrariness to this decision since it appears that the root defect is the same in both cases—failing to represent a dependency relation that exists. On the other hand, connecting proportionality to this root defect helps to make its underlying rationale more transparent.

With this as background, I suggest the following formulation of a proportionality requirement as a replacement for the formulation in Woodward (2010).

> (**P***) Suppose we are considering several different causal claims/explanations formulated in terms of different variables and representing different claims about patterns of dependency relations involving some target effect or explanandum *E* and where all of these satisfy some minimal interventionist condition like (**M***). Then, other things being equal, we should prefer those causal claims/explanations that more fully represent or exhibit those patterns of dependence that hold with respect to *E*.

The notion of representing or exhibiting should be understood along the lines described in Sect. 2. That is, (**P***) is to be understood as applying to choices among causal claims/explanations we are actually able to produce or exhibit. Thus (**P***) is better satisfied to the extent that causal claims and explanations are formulated in terms of variables and dependency relations that both fully capture those relationships on which *E* depends that do exist and do not imply relationships that do not exist. (**P***) should also be understood as applying to a fixed or pre-specified *E*. This specification, in addition to the empirical facts that obtain, fixes the range of possible variation that the effect or explanandum phenomenon (or explananda phenomena) exhibits and in turn what the causal claims/explanations we are assessing are required to account for. For example, in the various pigeon examples, *E* is specified by the variable *PECKS* which (we assume) can take just two possible values. As an empirical matter, the pigeons will peck in some circumstances and not others and it is this we are trying to account for. Obviously a candidate cause might do well in satisfying (**P***) with respect to one effect or explanandum *E* and less well with respect to some other explanandum *E**—this is one reason why in applying (**P***) we need to specify what explanandum or explananda we have in mind.[13]

Returning to the examples considered earlier in this section, I take it to be obvious that (**P***) judges (3.2) to be superior to (3.1). To apply (**P***) to the CYAN example (3.4), note that (3.4) both fails to convey the information that other shades of red besides scarlet will lead to pecking and also fails to convey the information that non-red colors besides cyan will lead to non-pecking. So (**P***) correctly judges (3.4) to be deficient in comparison with (3.2) even though (3.4) is true according to **M*** (and (**M****)). Note that if we restrict ourselves to the variable *V* in the CYAN example, (3.4) *does* fully convey the dependence of pecking on the values of that variable—the problem with (3.4) is instead that there is another variable *RED* that can be used to better capture dependen-

[13] Without some specification of a target explanandum or a class of these, the desideratum in (**P***) that more rather than less information about the factors on which *E* depends be described will be ill-defined. This introduces a kind of interest or goal relativity into (**P***) since the choice of target explananda will reflect in part the investigators goals or interests. But this sort of relativization seems an unavoidable feature of any theory of explanation.

cies that (3.4) misses. This illustrates the point that (**P***) is to be applied comparatively to alternative claims framed in terms of different variables in the cause position.

Turning next to Sober and Shapiro, we should be able to see that (**P***) is not subject to their counterexample since (**P***) does not say that to satisfy proportionality (or to come closer to satisfying proportionality) the functions describing causal relationships must be monotonic or 1–1. The function F does an entirely adequate job of representing the full range of dependency relations in Sober and Shapiro's example and only those relations and so satisfies (**P***). In other words, as far as (**P***) goes, there is nothing wrong with causal claims such that some variations in the values of a cause variable are *not* associated with differences in the values of effect variables in a 1–1 manner, as long as the actual patterns of dependence of the values of the effect on its causes are accurately and fully described.[14] (Of course if no variations in the cause variable are associated with any changes in the effect variable, then the cause is irrelevant to the effect.)

To enlarge on this, consider another variant on the pigeon case which will play a role in Sect. 5: a fine-grained color variable *G* takes distinct values for each of a number of distinct shades of red (scarlet, maroon etc.) and also different values for a large number of distinct shades of non-red colors, in this way covering the full color spectrum. This is accompanied by a specification of a functional relationship F which maps all the color values corresponding to the coarse-grained color red into the peck value of the *PECK* variable and all of the remaining values of *G* into the non-peck value of *PECK*. From the point of view of capturing the full range of dependency of *PECK* on target color, an appeal to

(3.5) F, along with fine-grained information about the target color presented on different occasions

seems to do as good a job as (3.2) in accounting for pecking behavior [and satisfies (**P***) just as well as (3.2)].

The most that might be said against (3.5) is that it employs a characterization of the cause-variable that is more fine-grained than is necessary given the effect-variable and in this respect is a less efficient or economical representation than (3.2). But this does not show that (3.5) fails to capture the full pattern of dependence of *PECKS* on target color or that it misrepresents that dependence—that is, that it fails to conform to (**P***). Although (3.5) employs a variable that involves discriminations among its values that are overly fine-grained or irrelevant as far as pecking goes, it "corrects" for this by specifying exactly which of those values leads to pecking and which do not, so that the resulting overall dependency relation is just what is conveyed by (3.2). On the other hand (to anticipate my discussion below) note also that for the example as described, if we are interested in accounting just for contrast between the pigeon's pecking versus not pecking, (3.2) and the *RED* vocabulary do just as good a job with respect to (**P***) as the more fine-grained vocabulary associated with *G*. As far as (**P***) goes, this is enough to entitle us to use (3.2) rather than the explanation that appeals to *G*. This illustrates how proportionality can *license* the use of coarse-grained variables, even if

[14] Assuming that the relationships with which we are dealing are deterministic, a necessary condition for satisfying (**P***) is that the function from the explanans to explanandum variables be onto.

it does not *require* this. In fact, as we shall see, much of the interest of proportionality lies in this fact—in its licensing or permissive role.

4 Lumping and proportionality

After suggesting that we might deal with her *CYAN* counterexample (3.4) to proportionality by supplementing this notion with the requirement that the variables employed "exhaust" the possibility space (so that a variable whose only values are "scarlet" and "cyan" is not a good variable because it is non-exhaustive, since there are other possible colors), Franklin-Hall advances an additional objection to a proportionality constraint. Suppose that in the pigeon example, in addition to presentation of a red target causing the pigeon to peck, tickling the pigeon's chin or electrically stimulating its cerebellum also causes pecking. Now consider

> (4.1) The presentation of a red target or provision of food or tickling of the chin or electrical stimulation of the cerebellum (other value: none of the above) causes the pigeon to peck (other value: not-peck).

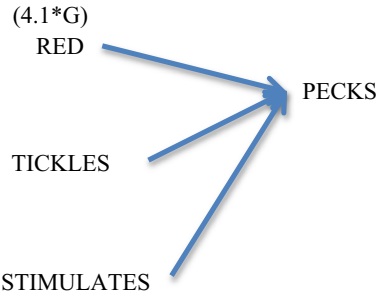
Franklin-Hall claims that (4.1) does an even better job of "exhausting the causal possibility space" than (3.2) (the *RED/PECKS* explanation) and should be judged superior to (3.2) by a plausible proportionality requirement. More specifically, she claims that (4.1) shows that proportionality combined with exhaustivity "recommends maximally disjunctive accounts, those citing causes that effectively lump together, into a single explanatory factor, every single means by which the effect might, in principle, have been brought about." (2016, p. 568). Franklin-Hall contends that "such [lumping] accounts are absent from the explanatory annals, presumably in part for their genuine explanatory inferiority; they are pitched at such great heights as to induce a kind of explanatory hypoxia, specifying far too little about what actually brought about the explanandum event to be very explanatory of it." (2016, p. 568)

In assessing this claim, we need to pay attention to a crucial distinction—the distinction between (different) *variables* and different *values* of those variables.[15] This distinction yields two different possible readings of (4.1). On one reading—call it (4.1*)—three distinct variables are described in the cause position of (4.1): a variable corresponding to whether the target is red, a variable corresponding to whether there is stimulation and a variable corresponding to tickling. On a second reading (4.1**) these variables are collapsed into a single "disjunctive" variable V^* with two possible values—"true" if either the target is red or there is stimulation or tickling and "false" otherwise. (4.1**) is then interpreted as claiming that V^* causes pecking. I believe this second interpretation is the one Franklin-Hall intends (it is the one that involves

[15] The distinction is important for many other reasons besides being required for a proper understanding of proportionality. I lack the space to discuss these here but see Woodward (2016a) for some additional applications. I also note that we need some account of when variables are distinct to capture the notion, discussed in Sect. 5, of the dimensionality or degrees of freedom associated with a model or explanation. For example, it is because the three variables representing the position of a particle in a three dimensional space are distinct and similarly for the three variables representing its momentum that there six degrees of freedom associated with the particle.

"lumping"), but, as we shall see, she also ascribes features to the example that are more consistent with the interpretation (4.1*).

It may seem tempting to suppose that there is no real difference between (4.1*) and (4.1**)—that both are equally good representations of the same causal structure and that in general it is a matter of indifference whether we employ a number of distinct variables in our causal representation or collapse these into a single variable, as (4.1**) does. One powerful reason for thinking otherwise is that the use of directed graphs to represent causal relationships and the methodologies for causal discovery associated with these depend crucially on there being such a distinction.[16] The graphical representation of (4.1*) is:



By contrast the graphical representation of (4.1**) is

Causal discovery algorithms like those described in Spirtes et al. (2000) depend on there being a difference between these two representations. (4.1*G) describes a collider structure which licenses certain inferences (e.g., that *RED* and *TICKLES* are dependent conditional on *PECKS*). (4.1** G) obviously does not imply this. Similarly, when S (in the smoking example) is represented as the common cause of the distinct variables Y and L, this allows us to apply the Causal Markov condition and conclude that Y and L are independent conditional on S. If we collapse Y and L into a single (presumably four-valued) variable X, we have S causing X and we can no longer apply the Causal Markov condition in this way.

Assuming that the choice of graphical structure to represent causal relationships is not completely arbitrary, there must be some basis for decisions about when it is preferable to represent a causal structure by means of distinct variables and when it is permissible (or a good strategy) to lump or collapse these into a single variable. I will not propose a general account of this but, following Hitchcock (2012) and Woodward (2016a), I take one relevant consideration to have to do with a distinction between variables and their values: A single variable can take any one of a number of different values for different systems or on different occasions but when a variable is predicated of a single unit or object, this variable cannot take two different values at the same time—e.g., a particular cannonball cannot have a mass of 10 kg and 20 kg at the

[16] A similar point holds for the use of structural equations to represent causal relationships.

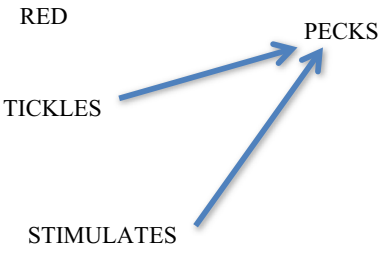
same time. (Here the force of "cannot" is logical or conceptual, rather than causal.) On the other hand, for systems of the sort described in Franklin-Hall's example, it is plausible that if two variables—call them V_1 and V_2—are fully distinct, all pairwise combinations of their values should be logically or conceptually possible (although there may of course be *causal* relationships among them that causally exclude certain possible combinations of values[17]). In other words, if V_1 takes the value v_{11}, this should not constrain, for logical or conceptual reasons, the value taken by some fully distinct variable V_2. (One might take this as a proposal for what it means for variables to be fully distinct.) Variables that are not distinct in this sense cannot stand in causal relationships with each other. For similar reasons, if one holds that variables can be experimentally manipulated independently of each other or that causal relationships involving these variables can be independently changed or interfered with, one is treating these variables as "distinct" in the sense just described.[18] Representations in which distinct variables are represented as such can thus allow us to capture facts about what would happen if causal relations in which those variables stand are independently disrupted (which involve answers to one kind of what-if-things-had-been-different question)—facts that may not be captured by alternative representations.

It seems clear that Franklin-Hall's description of (4.1) involves the assumption that it is possible for an experimenter (or nature) intervene to set the values of each of the cause variables independently of values of the others—e.g., by varying the color of the target from red to not red and, independently of this, deciding whether or not to tickle the pigeon. She thus treats these cause variables as distinct. Moreover, Franklin-Hall explicitly assumes that the "pathway" by which presentation of a red target causes pecking is distinct from the pathway by which other causes of pecking operate, and that because of this distinctness, these pathways may have independent conditions of breakdown and that it may be possible to interfere with one pathway independently of the other. To use Franklin-Hall's example, it may be possible to interfere with the *RED*→ *PECK* pathway by blindfolding the pigeon without disrupting the *TICKLES—>PECK* pathway. Moreover, whatever might be meant by "distinct pathway", it seems to be an uncontroversial necessary condition for pathway distinctness that the *RED* variable be distinct from the *TICKLES* variable in the sense described above. By contrast if we use a single variable *V** as in (4.1**) to represent the cause of pecking, we do not capture the facts about independent variability of values of and independence of pathways just described. That is, (4.1**) does not represent the fact that we (or nature) can alter the value of the variable *RED* independently of the value of the variable *TICKLES* (and what would happen under such an alteration) in the way that (4.1*) does. Thus the three-variable representation (4.1*) provides a better representation of how Franklin-Hall understands her example than (4.1**).

[17] This assumption—that all combinations of variables that are logically or conceptually possible—will be violated for systems in which non-causal dependency relations such as supervenience relations are present—see Woodward (2015a) for further discussion. But such relations are not present in Franklin-Hall's example (4.1).

[18] Although I lack space for discussion, the distinction between values and variables obviously has a additional implications for what it is for a predicate or property (or a cause) to be "disjunctive". Among other things, we need to distinguish between causes that act disjunctively (i.e., as an "or" gate) and causes or properties that have disjunctions as their values. We can't make sense of the notion of proportionality without something like the variable/value distinction.

Putting this in terms of graphical representations, (4.1*G) allows us to represent the result of interfering with the RED → PECK pathway while the other pathways remain intact as a matter of replacing 4.1*G with



This (arrow-removal) representation (and its non-graphical counterpart) correctly capture what would happen to *PECK* if we disrupt the *RED* → *PECK* pathway and simultaneously alter whether there is tickling and so on. By contrast (4.1**) and (4.1**G) do not represent this information. Thus in comparison with (4.1*), (4.1**) fails answer some what-if-things-had-been-different questions associated with this example and fails to represent dependency relations present in the example. One may think of this feature of (4.1**) as indicating that it satisfies proportionality (understood along the lines of **P***) less well than (4.1*). In other words (4.1*) is preferable to (4.1**) from the point of view of proportionality.

For these reasons, if the facts are as described by Franklin-Hall, I see nothing problematic about (4.1) when understood as (4.1*)–(4.1*G). Instead, the judgment implied by the w-question criterion that (4.1*) is preferable to (4.1) seems correct: (4.1*) conveys a wider range of information about what *PECKS* depends on. Moreover, this judgment is reflected in aspects of scientific practice. For example, if we wanted to construct a regression model of the pigeon's pecking behavior, it would generally be thought preferable to construct a multivariate equation which corresponds to the graphical representation (4.1*G) with the variables all explicitly related to the dependent variable *PECKS* rather than a bivariate regression equation with a single independent variable.[19] There are legitimate questions in such cases about the criteria for when one should stop adding independent variables to a regression equation (see "Appendix 2") but it is not always regarded as methodologically objectionable to add them. Once we add requirements concerning the conditions under which variables are distinct, we see that neither (**P***) nor the w-question criterion license an indiscriminate collapsing of distinct variables into a single general variable, with an accompanying "explanatory hypoxia".

[19] Again, recall that the example is construed as a type-level causal claim. Of course on any given occasion there presumably will be a fact of the matter about whether a particular episode of pecking is caused by the presentation or a red target or by tickling or by some combination of these. If a particular episode of pecking, e, is caused by the presentation of a red target and nothing else, then the fact that the pigeon would have pecked if tickled is arguably explanatorily irrelevant to e. An account of actual causation like that in Woodward (2003) will yield this conclusion. The *TICKLES* → *PECK* relationship does become relevant if we are interested in a type-level explanation of pecking behavior.

5 Autonomy and conditional irrelevance

In this section, I want to use some of the ideas from previous sections to explore some larger issues about the status of upper-level explanations and how considerations having to do with proportionality and stability can help us in finding the right level or variables for framing such explanations.

Let me begin by being a bit more explicit about the ways in which considerations of proportionality can guide variable choice. First, a cause or explanans variable may be such that its values cannot be used to provide a full accounting of the range of conditions under which the explanandum takes its values and some alternative candidate for the cause variable may do better on this score in which it case it should be preferred. This is the reason for preferring *RED* over *SCARLET* in the pigeon example. A second possibility is more subtle and has to do with the way in which in certain situations considerations having to do with (**P***) can license the use of more coarse-grained or macroscopic variables over more fine-grained microscopic variables.

To motivate this idea, I begin with a striking empirical fact: to an extent that may initially seem surprising, the difference-making features cited in many lower-level, fundamental theories sometimes can be absorbed into variables that figure in upper-level theories without a significant loss of difference-making information with respect to the explananda of those upper-level theories. This fact is crucial for understanding how upper-level explanation is possible. To capture this, I begin with the simplest possibility, which I acknowledge is a limiting case, and then relax some of its characterizing assumptions. Suppose it is possible to find a set of variables X_i which are causally relevant to explanandum E holding for system S and which are such that, given the values of those X_i, further variations in some other set of variables Y_k characterizing S are irrelevant to (do not make a difference for) E, even though the Y_k have much higher dimensionality or degrees of freedom than the X_i. In the example from Sect. 3, given the values of the *RED* variable, further variations in the values of the more fine-grained color variable G in 3.5 are irrelevant to whether the values taken by the *PECKS* variable. As another illustration suppose (as is almost but not quite true) that variations in the values of various thermodynamic variables like temperature (the X_i variables above) are difference-makers for those aspects of the macroscopic behavior of a gas that are described by thermodynamic variables (E), and that further variations in its microscopic state (the Y_k variables above), as described by the positions and momenta of the individual molecules making it up that are consistent with the values taken by the macroscopic variables are irrelevant to the behavior of the gas.

To further spell this out, let us say, following (**M***), that a set of variables X_i is *unconditionally relevant* (alternatively, irrelevant or independent) to E if there are some (no) changes in the values of each X_i when produced by interventions that are associated with changes in E. A set of variables Y_k is irrelevant to variable E *conditional* on additional variables X_i if the X_i are unconditionally relevant to E, the Y_k are unconditionally relevant to E, *and* conditional on the values of X_i, changes in the value of Y_k produced by interventions and consistent with these values for X_i

are (unconditionally) irrelevant to E.[20] In other words, changes in the variables X_i are causally relevant to E in the sense captured by (**M***) and conditional on the values taken by X_i, further variations in the Y_k make no difference to E.[21] We can think of this as a generalization of the "screening-off" idea used by Yablo to characterize proportionality described in footnote 11—the X_i screen off the Y_k from E.[22]

In a case of this sort, as noted in Sect. 3, we can satisfy the demands associated with (**P***) just as well by citing the X_i to explain E as by citing Y_k: Thus proportionality understood as (**P***) does not drive us in the direction of always preferring more fine-grained or microscopic variables: if we want to explain E, we can just cite the variables X_i, and ignore the lower level variables Y_k. As it is often put, one collapses the many degrees of freedom or the high dimensionality associated with variations in Y_k into the much smaller number of degrees of freedom associated with X_i. Put differently, the X_i are a permissible coarse-graining of the Y_k with respect to E. Thus one does not *need* to model the system in terms of the Y_k, the X_i do just as well. And of course, this is a very good thing if, as a practical matter, there is no possibility of actually

[20] Two observations: First, I want to underscore that relevance/irrelevance are understood in terms of counterfactuals describing what happens under interventions (rather than statistical dependence)—e.g., if X is conditionally irrelevant to Y, given Z, then, if (1) one intervenes to fix Z at some value, (2) further variations in X due to interventions consistent with (1) will not change Y. Second, conditional irrelevance is much stronger than multiple realizability. The latter requires only that *some* different values of the same or different micro-variable(s) realize the same value of a macro-variable. Conditional irrelevance requires that *all* variations at the micro-level consistent with the value of the macro-variables make no difference to E. As this observation suggests, multiple realizability is *not* sufficient for autonomy understood in terms of conditional irrelevance.

[21] An anonymous referee notes that according to my definition, conditional irrelevance is not a special case of unconditional irrelevance, in the way in which such a relationship between conditional and unconditional irrelevance holds in probability theory, as when unconditional irrelevance is viewed as a special case of conditioning on a tautology. The referee writes, "[s]imilarly, one might expect that unconditional irrelevance in the present case [that is, in the applications I have discussed] comes out as a special case of conditional irrelevance where we conditionalize on an empty set of variables. However, …it seems to me that this is not the case." This is an interesting observation/suggestion for which I thank the referee. The referee is correct that on the characterizations given above unconditional irrelevance is not a special case of conditional irrelevance. However for several reasons this does not seem to be a defect in my characterization (or contrary to what we should expect). First, recall that my characterizations of both unconditional and conditional relevance/irrelevance are understood in this paper in terms of interventionist counterfactuals rather than in terms of notions of (in)dependence and conditional (in)dependence as these are understood in probability theory. For this reason alone it is not clear that we should expect the relation between unconditional and conditional irrelevance as I have characterized these notions to behave like the relation between their probabilistic counterparts. Second, for the purposes of this paper I am only interested in cases in which one conditionalizes via interventions on a non-empty set of variables—nothing in the paper requires me to take a position on how we might understand the notion of intervening on an empty set of variables or even whether this notion (or notions that might be defined in terms of it) makes sense. Third, if one wants to preserve the analogy with probabilistic independence, an alternative response might be to drop the requirement in the definition of conditional irrelevance that the Y_k be unconditionally relevant to E, allowing Y_k to be irrelevant to E conditional on X_i even if Y_k is unconditionally irrelevant to E. One could then propose identifying the unconditional irrelevance of Y_k to E with the irrelevance Y_k to E conditional on the empty set. However, this has the disadvantage of not focusing on the kind of example which the characterization in the text is intended to capture, in which the Y_k are (unconditionally) relevant variables whose relevance to E is fully absorbed by the X_i.

[22] Note that, as with **(P*),** relevance and irrelevance are always defined relative to an effect or explanandum E. Y may be irrelevant to E conditional on X but this may not be true for some alternative explanandum E^*.

constructing or exhibiting explanations that appeal to the variables Y_k; and associated laws, but one can exhibit explanations that appeal to X_i.

As I have said, this is an ideal case but I take it to provide an illustration of how and in what circumstances upper-level explanation is possible and how such explanation can be "autonomous" from lower-level details—autonomy here just means that the upper-level variables are relevant to the explanandum E and that the variables figuring in lower-level or more fine—grained theories are conditionally irrelevant to E given the values of the upper-level variables.[23]

This framework differs from the vindication of the non-pragmatic superiority of upper level explanations sought by Weslake and Franklin-Hall in several important ways. First, there is no attempt to argue that the explanation of E in terms of the upper-level variables X_i is "better" than the explanation in terms of the lower-level variables Y_k given the contrary-to-fact supposition that one can exhibit the latter explanation. Rather, the idea is that the explanation of E in terms of X_i is *no worse* than the explanation in terms of Y_k even assuming we were able to construct the latter. Second, there is a clear sense in which this justification for the use of X_i is not "purely pragmatic". Pragmatics does play some role in the justification since in applying **(P*)** we choose among explanations we are able to formulate and this will reflect facts about our limitations.[24] However, given that our choice is restricted in this way, pragmatics (in the sense under discussion) does not enter into the evaluation of these explanations in any further way that is disturbing. Thus although our limitations do play a role in our decision to employ the upper-level explanation, the fact that the upper-level explanation is no worse (and that the appropriate conditional irrelevance relations hold) is not just a matter of pragmatics—this fact depends on what the world is like and its significance for explanation depends on the adoption of an account of explanation in which con-

[23] Some brief remarks about similar ideas found elsewhere in the literature are appropriate here. First, Woodward (2008) describes the notion of a *realization independent dependency relationship* (RIDR). This involves a dependency relationship between upper- level variables M_1 and M_2 that continues to hold for a range of different lower- level realizers for M_1 and M_2—that is, interventions that change M_1 (and involve different lower-level realizations of the same value of M_1) are stably associated with changes in M_2 (also involving different lower-level realizations of the same value of M_2). The notion of conditional irrelevance introduced above attempts to capture the same basic idea but in a way that is more general and (I hope) somewhat more precise.

Second, in several papers (e.g., 2009, 2010), List and Menzies introduce a very similar notion involving what they call "realization-insensitive causal relations"-- these are upper-level causal relations that are invariant under perturbations of their lower-level realizers. They argue that when there is realization-insensitivity, the appropriate level of causal explanation is the upper one; when there is realization-sensitivity, it is not. List and Menzies also give precise formal definitions of these notions. Their underlying conception of what is required for autonomy is in a number of respects very similar to the notion defended above—the underlying conception is that autonomy involves a kind of insensitivity to (or independence from) lower level details, given a specification of upper-level variables. However, my treatment differs from theirs in that I do not claim, as they do, that the truth of the upper level causal claim "excludes" the truth of the lower-level causal claim. Relatedly, I follow Woodward (2017a) in not assuming, as List and Menzies do, that the satisfaction of a proportionality condition is necessary for the truth of causal claims.

Finally, rather similar ideas are developed in considerably more formal detail in Chalupka et al. (2015, 2017).

[24] It is also true, as noted earlier, that we choose among explanations that are directed at explananda we are interested in explaining, a consideration that might also be regarded as "pragmatic". However, any theory of explanation will need to acknowledge this feature.

ditional relevance and irrelevance play the role just described.[25] In this connection, we should also note that, on the account proposed, the invocation of "pragmatics" by itself is not sufficient to explain why upper-level explanations are sometimes possible. In some cases very high dimensional lower-level variables may be relevant to some upper-level explananda E and there may not exist more coarse-grained upper-level variables (of a sort that we are able to formulate or measure) that satisfy the screening-off requirements just described. In such cases, actually exhibiting an explanation of E may be impossible, however strong our pragmatic reasons for wanting to do this may be. This is the "model chaos" described by Goldenfeld and Kadanoff below. What we want in an account of upper-level explanation is not just a role for pragmatics, but also (1) some insight into conditions in the world that sometimes support or allow for the successful construction of upper-level explanations as well as (2) an account of explanation that illuminates how such explanations work. My claim is that this has to do with the obtaining of the right sort of conditional irrelevance relations.

I acknowledged above that the sort of complete conditional irrelevance of *all* lower-level detail from an upper-level explanandum just described may not be common, although it may not be highly uncommon once we restrict its deployment to explanations and relationships we are able to exhibit. We can weaken the notion of conditional irrelevance in various ways. One possibility is that although there may be rare or exceptional values of the Y_k that are conditionally relevant to E, even given the values of X_i, this may not be true for most or "almost all" values of the Y_k—for most or almost all such values, the Y_k are conditionally independent of or irrelevant to E, given X_i even if there are a few Y_k for which this is not true. Or perhaps conditional irrelevance holds for all values of the Y_k and X_i within a certain large interval, including those values most likely to occur (at least around here right now). In such cases, standard explanatory practice often is to explain E just by citing the X_i, again especially if it is impossible to actually construct or exhibit an explanation of E in terms of the Y_k. Again, we can think of this as providing a justification for the use of the upper-level explanations that is not purely pragmatic—the justification also has to do with the facts about "almost irrelevance" just described.[26]

[25] Of course a lot depends on what is meant by pragmatics (and by "purely pragmatic"). Suppose that I employ criteria for hypothesis choice that are (let us stipulate) not pragmatic in any sense but use these to choose among hypotheses that it is possible to exhibit, which we stipulate reflects human limitations -- a pragmatic consideration. Does it follow that the result is a "purely pragmatic" account of hypothesis choice? This seems like a misleading or unnuanced way of characterizing the situation. Better to recognize that "pragmatic" considerations can enter into assessments in many different ways and that we should discriminate among these, rather than lumping them all together.

[26] I acknowledge that this is a point at which considerations that are pragmatic in the sense of reflecting cost/benefit considerations *may* enter the picture. Some departures from full conditional independence may reflect the influence of factors that are so small or rare that it is thought not to be worth it to complicate a model by including them. However several additional points about this are worth noting. First, the smallness or rarity of the omitted factors is not *just* a matter of pragmatics—it also reflects what nature is like. Second, that cost/benefit considerations enter in this way does not (in my view) show that there is anything wrong with the claim that it is a consideration in favor of an explanation that it answers more w-questions rather than fewer over a large range of such answers. It just shows that something else ("cost") matters in addition to answering w-questions. Finally, I emphasize again that in real-life scientific explanations, it is often *not* the cost or complexity of including additional factors that leads us not to introduce them but rather the *impossibility* (because of calculational and other limitations) of doing so in a way that exhibits the dependence of the explanandum on these factors.

Here are some examples. I noted above that to the extent that our target explananda involve thermodynamic variables describing the macroscopic behavior of a gas like temperature, pressure and volume, it is almost but not quite true that microscopic variations consistent with the values taken by these thermodynamic variables are irrelevant. What is more nearly true is that this is so for almost all—in the sense of all but a set of Lebesgue measure zero—values taken by those microscopic variables. Thus given this target, we can replace the enormous number of variables (and degrees of freedom) necessary to characterize the full microscopic state of the gas with a much smaller number of variables while still satisfying a proportionality requirement like (**P***) as well as making use of stable relationships that bear on what we want to explain.[27]

As a second illustration, consider the following remarks of Goldenfeld and Kadanoff (1999) concerning a simple computational model that reproduces real features of fluid flow despite omitting most details concerning the micro-behavior of the constituents of the fluid. They write:

> For physicists it is delightful, but not surprising, that the computer generates realistic fluid behavior, regardless of the precise details of how we do the coding. If this were not the case, then we would have extreme sensitivity to the microscopic modeling—what one might loosely call "model chaos"—and physics as a science could not exist: In order to model a bulldozer, we would need to be careful to model its constituent quarks! Nature has been kind enough to have provided us with a convenient separation of length, energy, and time scales, allowing us to excavate physical laws from well-defined strata, even though the consequences of these laws are very complex (87)

Again, the point is that for many aspects of the systems of interest, variations in microscopic variables either don't matter at all, given the values taken by certain macroscopic variables, or matter only in certain unusual cases. This is very fortunate since it makes it possible to model or explain important aspects of the behavior of those systems without adverting to these microscopic details. If we had to appeal to these details (which would be the case if even near conditional irrelevance relations fail to hold), the exhibition of an explanatory model would be hopeless. As Goldenfeld and Kadanoff suggest, this possibility (of neglecting various low-level details) is closely linked to the physical fact of the "separation of scales" which in an interesting range of cases, has the result that phenomena occurring at length, time and energy scales S are largely or entirely conditionally independent of what is going on at other scales S* with more degrees of freedom. If we can actually construct (display) a model that explains a range of explananda at scale S in terms of some coarser grained theory T and we cannot do this in terms of some finer-grained theory T* (which, if we could construct it, would also allow for the explanation of phenomena at scale S*) and conditional irrelevance relations of the form described above hold, then T satisfies (**P***) and the w-question criterion with respect to these explananda.

Goldenfeld's and Kadanoff's view is very similar to the view recommended in this paper. They don't try to argue that an upper-level explanation of bulldozer behavior

[27] Other examples include various forms of "universal behavior" exhibited by materials that differ greatly in the micro-details, as discussed in a series of papers by Batterman (e.g., 2000).

would be superior to one in that appeals to quantum chromodynamics even if we could derive bulldozer behavior from this theory. Instead they argue that we don't need to appeal to QCD because of facts about what nature is like (separation of scales) and this is very fortunate, since modeling in terms of QCD (in the sense of actually exhibiting such a model) is impossible.

We may compare Goldenfeld's and Kadanoff's remarks with the following claim from Franklin-Hall:

> … judged by [an interventionist] standard of excellence, high-level explanations are uniformly impoverished; they explicitly represent fewer features of the world on which the explanandum depends than do lower–level "micro" explanations, limiting the range of w-questions they can answer. (2016, p. 554)

Even putting aside the point that often it is not true that lower-level explanations "can answer" w-questions about upper-level phenomena in the sense of actually exhibiting such answers, this claim of "uniform impoverishment" seems an exaggeration when "depends" is interpreted in terms of conditional dependence, which I suggest is the most obviously relevant interpretation. It is true that the micro-explanation of, e.g., the behavior of a gas contains a *representation* of many features not represented in the macro-explanation, but the crucial point is that the macroscopic features of the gas that we are interested in explaining may not *depend* (or may not depend in almost all circumstances) on these microscopic features, conditional on the other macroscopic variables characterizing the gas. Franklin-Hall's discussion does not adequately reflect the role of this fact in allowing us to (justifiably) omit reference to these features in constructing an explanation of the macroscopic behavior of the gas.[28]

I acknowledged above that on the account I have defended many upper-level explanations are not fully autonomous. Some readers may find this disturbing or at least disappointing—they were perhaps hoping for some stronger result. I disagree: Insofar as there is some feature of good explanatory practice that needs to be captured or explained, the feature in question is *not* that (1) upper level explanations that eschew appeal to all lower-level considerations are always or even usually superior to those that do, so that a non-pragmatic account of upper-level explanations must vindicate or support (1). Instead, a non-pragmatic account of explanation that implies (1) is misguided for the reason just described. Here the tendency to identify "lower-level" with fundamental physics (or to think in terms of a simple dichotomy between "upper' and "lower" level) misleads us. It may well be true, for example, that explanations of behavior appealing only to "psychological" variables are unlikely to be improved by incorporation of information from fundamental physics (in part but not only in part because we have no idea how to do this), but it is plausible that *neurobiological* variables are relevant to many psychological phenomena even conditional on psychological variables. If so, an explanation that incorporates neurobiological variables as well as psychological variables in a mixed level explanation will be superior to one

[28] Suppose, on the other hand, that we are in a situation in which even approximate conditional irrelevance fails for our upper-level theory T with respect to explanandum E and we have model chaos. Then it *will* be true that there are features of the world on which E depends that are not represented in T. I do not understand, however, why, as Franklin-Hall seems to imply, this is a problem *for interventionism.* Instead, interventionism correctly judges that T is explanatorily inadequate.

that appeals only to psychological variables. This seems to me the correct assessment. We shouldn't be looking for an account that implies otherwise.

6 Weslake on non-pragmatic superiority

I noted above that Weslake advances a positive proposal about what the non-pragmatic superiority of "upper level" explanations consists in. Very briefly, he holds this is a matter of there being "physically impossible [but metaphysically possible] systems to which the macroscopic explanation applies but to which the microscopic explanation does not" (2010, p. 287) so that in this sense the former is "more abstract" and applies to a wider range of possibilities than the former. The presence of these features can make the upper level explanation non-pragmatically better. For example, Weslake claims that the ideal gas law would hold in a world in which the underlying mechanics is Newtonian in addition to the actual world which is quantum–mechanical (2010, p. 291). According to Weslake, this makes an explanation in terms of the ideal gas law non-pragmatically better than a quantum–mechanical explanation.

I do not find this convincing. To begin with, the fact that there are metaphysically (or logically) possible but physically impossible scenarios in which the ideal gas law holds seems in itself irrelevant to its explanatory goodness. Consider a contra-nomic but arguably metaphysically possible scenario in which little men move molecules around in a way that conforms to the ideal gas law.[29] Why should this possibility contribute anything to the "depth" of the explanations the ideal gas law provides? Or suppose that there are (in the relevant sense) possible but non-existent systems (e.g., composed of silicon) for which the generalizations of folk psychology hold exactly. Why should this fact be relevant to the assessment of the explanatory depth of folk psychology as applied to human beings?

The considerations to which Weslake appeals in connection with the ideal gas law are better captured in the following way, which appeals to the ideas about conditional irrelevance described above. It is a mathematical fact that any underlying micro-theory having certain generic features will lead to the ideal gas law. The quantum mechanical theory that correctly characterizes the actual world has these generic features. As it happens, these generic features or something close to them are also shared by certain quasi-Newtonian models of the gas. Put in terms of the framework described above, we can thus appeal to these generic features rather than more specific details of quantum mechanics to explain the behavior of gases—given the generic features, the ideal gas law is conditionally independent of these more specific details. However—and this is the crucial point—it is the fact that these generic features hold for the actual quantum mechanical laws governing our world that establishes their explanatory relevance to gas behavior. The Newtonian models "inherit" their explanatory relevance from the features they possess that are shared with actual quantum mechanical laws and not because there is a general preference for explanations whose scope covers contra-nomic worlds. Weslake mistakenly interprets the correct idea that whether the ideal gas

[29] If it is claimed that this scenario is not in the relevant sense "possible" we are owed an explanation for why this is so, which Weslake does not provide.

law holds is conditionally independent of certain details of the underlying physics (and that this has implications for its explanatory status) as a claim about the explanatory relevance of the law's holding in physically impossible situations. As I have tried to illustrate, conditional independence claims need not be understood in this way.

If, as I have argued, Weslake's attempt to show that "upper level" explanations are sometimes non-pragmatically better is unsuccessful, then unless there is some other reason for accepting this claim about non-pragmatic superiority, this provides an additional reason for not demanding that accounts of upper-level explanation establish this claim.

I can further clarify the relationship between my own proposal and Weslake's by commenting briefly on Weslake's discussion of Woodward (2003) on the relationship between macro-level and micro-level explanations of the behavior of an ideal gas. Weslake takes Woodward to be claiming that there is modal information in the former that is not provided by the latter and that, even abstracting away from our limitations, the former answers w-questions not answered by the latter. He rejects these claims, writing

> If we assume a reasonable form of physicalism, then there are no questions that can be formulated in terms of any other variables that do not correspond to one of these questions [about the values of microscopic variables]. So there are no physically possible counterfactuals on which the fundamental physical explanation is silent. The fundamental physical explanation provides the resources to answer any possible w- question. …there is no missing modal information of the kind claimed. (281)

I agree with Weslake that there is no missing modal information. However, I don't think this observation has the significance that Weslake takes it to have. In particular, it does not undermine the arguments advanced above about role of conditional independence in licensing more upper-level explanations of the behavior of the gas. In particular, insofar as it is possible to extract claims about the macroscopic behavior of the gas from the underlying microphysics, these will themselves show that certain variations in the underlying microstates of the gas are irrelevant to its macroscopic behavior and hence can be absorbed into a macroscopic representation with fewer degrees of freedom. Put differently, the claim that the macroscopic behavior of the gas is "implicit" in the underlying microphysics does not distinguish between two very different possibilities: that (1) the underlying physics shows that variations in those micro-details are relevant to macroscopic behavior in a way that cannot be captured by a few macro-variables and that (2) the underlying physics shows that the macroscopic behavior can be so captured by such macro-variables. In the former case, but not the latter, we have Goldenfeld's and Kadanoff's model chaos, in which to model the macroscopic behavior we must model the miro-constituents in full detail. Both possibilities (1) and (2) are consistent with Weslake's implicitness claim and his claim about what physicalism requires, but have very different implications for the possibility of upper-level explanation that neglects micro-details.

7 Conclusion

I have, in effect, urged that we should ask a different question about upper-level explanations than the question that animates Weslake and Franklin-Hall. Rather than asking, as they do, whether, under interventionist assumptions, abstracting away from epistemic and calculational limitations, upper-level explanations would still be superior to fundamental explanations, a better question is this: what features of the world and what conception of explanation make it possible, given our limitations, to sometimes formulate successful upper-level explanations?

Acknowledgements I would like to thank Thomas Blanchard and Stephen Yablo for very helpful comments on an earlier draft of this paper.

Appendix 1: More on non-pragmatic superiority

Both Weslake and Franklin-Hall argue that some upper-level explanations are non-pragmatically superior to explanations framed in terms of fundamental physics and criticize interventionism for implying the contrary conclusion. Part of my response to this criticism is that the interventionist criteria for explanatory assessment (proportionality, stability, and the w-condition criterion) are meant to apply only to explanations that are actually produced or exhibited. The interventionist criteria are intended as contributions to methodology in the sense of Woodward (2015b) and methodology, as I see it, has to do with choices among possibilities that are available or realistically possible.

Putting this consideration aside, there are other reasons why we should be skeptical of Weslake's and Franklin-Hall's arguments. First, it is unclear why we should attach much (if any) weight to intuitive judgments about the non-pragmatic superiority of explanations appealing to upper-level theories in comparison with explanations of the same explananda in terms of fundamental physics, under the counterfactual assumption that we are somehow able to construct the latter. I, for one, have no strong "intuition" about whether, say, explanations of the behavior of financial markets in terms of economic and financial variables would be (non-pragmatically) better than explanations of that behavior in terms of the standard model of physics, given the fantastic hypothetical that we are able to produce the latter, in part because I have no clear conception of what this would involve. For those who have such intuitions, I ask why we should trust them. There does not seem to be anything in scientific practice that might serve as a guide to whether we are judging non-pragmatic merits correctly in the sort of case envisioned. But unless this intuitive judgment is correct, there is no basis for criticizing interventionism for failing to imply it. It is much better practice to assess interventionism (and proportionality and stability) in terms of what they imply about comparisons of explanations we are able to exhibit.[30]

[30] Woodward (2008, 2010) did not attempt to use proportionality and stability to compare upper-level explanations with potential explanations provided by fundamental physics. Instead Woodward [as well as scientists who have appealed to similar ideas (e.g., Kendler 2005)] attempted to use these considerations as a partial basis for choosing among different explanations, all of which are "upper-level" and non-fundamental.

Second, the argument under consideration "proves" too much. If the argument is cogent, it can be used to reject many other plausible criteria for explanation assessment. Consider criteria according to which explanations that appeal to fewer free parameters or have fewer degrees of freedom or that predict better are, ceteris paribus, superior to explanations that score less well according to these criteria. Suppose that abstracting away from the fact that we are not able to produce them, explanations of upper-level explananda in terms of fundamental physics do better in terms of these criteria than upper-level explanations of these same explananda. It would then follow, by the argument described above, that we have reason to reject these criteria as well, even when they are used to compare explanations that we are actually able to produce. Again, it seems much more plausible to conclude instead that we should not try to assess criteria which can be used to compare upper-level explanations that we are able to produce by considering what they imply about supposed intuitions regarding comparisons of non-pragmatic superiority concerning explanations we are not able to produce.

Appendix 2: More on the w-question criterion

The w-question criterion connects the goodness or depth of an explanation to its ability to answer a range of w-questions about an explanandum, as discussed in Woodward (2003) and Hitchcock and Woodward (2003). In addition to their more specific objections to Woodward's formulations of proportionality and his use of stability, one way of putting some of Weslake's and Franklin-Hall's more general criticisms is that the w-condition criterion lacks, as it were, a natural stopping point—they worry that it licenses the conclusion that more lower-level detail and more information about the causes that affect some outcome, however minutely, are always better, contrary to what they suppose is ordinary explanatory practice. Some of my responses to this criticism are given in the main text of this article—for example, the fact that we can't construct answers to w-questions from certain premises because of computational or epistemic limitations provides one natural stopping point and one that is crucially important in scientific practice. However, there is more that can be said. (Although part of what makes this complicated is that different things need to be said about different cases.)

First, one important constraint on what should be included in an explanation comes from the specification of the target explananda. In particular, the target explananda for scientific theories are typically repeatable phenomena or regularities rather than

Footnote 30 continued
For example, proportionality can guide us in choosing between explanations that appeal to neuronal firing rates and explanations that appeal to more detailed facts about neuronal behavior such as the time courses of firing. In my view, even if it is true that interventionism combined with proportionality and stability leads to the conclusion that no upper-level explanations are non-pragmatically better than fundamental explanations and even if this conclusion is "wrong", it does not follow that proportionality and stability cannot be legitimately used to choose among non-fundamental explanations. This is enough to show the philosophical importance of proportionality, stability, and the interventionist framework.

particular events in all of their complexity.[31] The w-question criterion was designed primarily to apply to such repeatable explananda. For example, in an example discussed in Woodward (2003), Coulomb's law is used to explain why the electrical field due to the charge distribution on a long straight wire takes a certain form. In the case of any actual wire, the actual field likely will be considerably more complex since it will reflect the influence of whatever other field sources are present in the vicinity, various inhomogeneities in the wire and so on. Thus if the goal is to explain the character of some actual field in all of its particularity, a hugely complicated explanans may be required, with lots of piling up of detail and perhaps no natural stopping point—there may always be some additional tiny effect that might be included. In practice, much of this complexity is avoided by taking the target explanandum to be that just that portion of the field which is due to the charge along the wire and how this changes under changes in the configuration of the wire. This is a repeatable phenomenon and taking it as the explanatory target allows us to dispense with non-shared detail that is idiosyncratic to particular cases. That we do not see an endless piling on of detail or (at least typically) citing of extremely long lists of causes in explanatory practice in much of science is thus in part a reflection of the sorts of explananda we try to explain. In the unlikely event that we did have the goal of explaining the field in as much particularity as possible, it is not obvious to me that a theory of explanation should imply that it is wrong to pile as much detail as possible into the explanans.

That said, there certainly are cases in which it is of scientific interest to explain particular outcomes (e.g., the Challenger explosion) or at least patterns that are highly concrete (what are the causes that influence student performance in U.S. public schools in 2017?). In such cases, given some effect or explanandum *E,* we often *select* some causes of *E* but not other causes of *E* to put in explanations or causal claims. We do so on the basis of a number of criteria, some of which are certainly "pragmatic"—for example, in the case of the Challenger, a quasi-normative consideration, having to do with the failure of the O-rings to behave as they were designed to behave may be crucial in selecting this factor as the cause. *If* the goal is to describe such *selection practices*, the w-question criterion may be of limited usefulness—the criterion was designed to compare explanations, not to describe causal selection practices.[32]

In the school case, a standard causal modeling approach will cite a number of different causes (student demography, training of teachers and administrators, level of financial support) but there are various natural criteria for stopping—for example, at some point the coefficients on additional variables that might be included will not be reliably statistically distinguishable from zero. The w-question criterion is not inconsistent with this practice.

[31] Both Franklin-Hall and Weslake focus on examples in which the explanandum is a particular event rather than a regularity or phenomenon. This focus has a major impact on their discussion since the explanation of particular events can be readily understood as having an open ended character with no natural stopping point—as Hempel (1965) observed, a particular event can be understood as indefinitely detailed and as calling for a similarly detailed explanation. The issues they discuss look quite different when one considers explanations of regularities.

[32] When we engage in causal selection, as in the Challenger example, we select one or some small number of causes from the very large number of factors that are causally relevant to some outcome. In this sort of case, we need not think that an explanation that cites only the O-rings is "better" (in some non-pragmatic$_1$ sense) than one that cites the O-rings and other causal factors as well, even if the former is the usual practice.

Finally, let me return to an observation made in Sect. 5—that relevance and irrelevance (as well as autonomy) must be understood as relative to some effect or class of effects E. In virtually all real-life cases what we find is that certain variables Y_k are conditionally irrelevant to some set of explananda E conditional on other variables X_i but that there are other explananda E^* for which this not true and which require the Y_k for their explanation. For example, thermodynamic variables render quantum mechanical variables characterizing the component molecules of a gas irrelevant to many behaviors of the gas but not to all—we need quantum mechanics to explain the specific heats of gases. Note, however, that once considerations having to do with the importance of actually exhibiting explanations are taken into account, it does *not* automatically follow that the Y_k will answer more w-questions that the X_i. Instead, what often happens in real life cases is that the Y_k can be used to answer w-questions about E^* but not about E and the X_i can be used to answer w-questions about E but not about E^*. So we have a set of different theories or models framed in terms of different variables each with its own proprietary set of explananda. As an illustration, consider a review paper (Herz et al. 2006, cf. Woodward 2017b) on neural modeling at different levels. A successful "circuit level" explanation of the behavior of an individual neuron, such as the Hodgkin–Huxley (HH) model, explains a range of different explananda by answering w-questions about them—it identifies the conditions under which an action potential will be generated (or not), how the shape of the action potential is affected by the cross membrane voltage and capacitance and so on. Of course there are many other questions about aspects of neuronal behavior this model does not address. For example, the action potential involves the opening and closing of individual ion channels in the neural membrane and the HH model does not tell us anything about the molecular mechanisms underlying these. However, as the authors explain, it is also not true that one can actually exhibit explanations of the circuit level behavior based only on molecular level variables—among other considerations this is a computational impossibility. So what one ends up with is a hierarchy of different models at different "levels" (the authors describe five such levels) each of which is capable of accounting for (actually answering w-questions about) some explananda and not others.[33]

References

Anderson, P. (2011). *More and different: Notes from a thoughtful curmudgeon*. Singapore: World Scientific.
Batterman, R. (2000). Multiple realizability and universality. *The British Journal for the Philosophy of Science, 51,* 115–145.

[33] Ironically, this separation of theories or models into levels, with (at least to a large extent) a proprietary set of explananda associated with each level is, if anything, even more true of fundamental physics. Here what can be actually calculated or solved, either analytically or by means of perturbation methods, is much more limited than many philosophers seem to realize. In general in theories like QED and QCD most of what can be calculated has to do with correlation functions for field values at various spacetime points from which information about scattering matrices can be extracted. It is this information that is used to test these theories. In the case of QCD, one cannot even calculate essential properties of protons and neutrons from the quark and gluon fields because there is no small parameter that can be used for a perturbative expansion at those relatively low energies. QCD can capture, e.g. proton/proton interactions at very short distances but not for longer distances. Explaining the properties of, say, a heavy nucleus, much less an atom, requires a different set of theories or models.

Blanchard, T. (forthcoming). Explanatory abstraction and the goldilocks problem: Interventionism gets things just right. *British Journal for the Philosophy of Science*.
Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: An overview. *Behaviormetrika, 44,* 137–164.
Chalupka, K., Perona, P., & Eberhardt, F. (2015). Visual causal feature learning. In *Proceedings of the thirty first conference on uncertainty in artificial intelligence*. AUAI Press, Corvallis (pp. 181–190).
Franklin-Hall, L. (2016). High level explanation and the interventionist's 'variables problem'. *British Journal for the Philosophy of Science, 67*(2), 553–577.
Goldenfeld, N., & Kadanoff, L. (1999). Simple lessons from complexity. *Science, 284,* 87–89.
Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. Free Press: New York.
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.
Hitchcock, C. (2012). Events and times: A case study in means-ends metaphysics. *Philosophical Studies, 160,* 79–96.
Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, Part II: Plumbing explanatory depth. *Nous., 37,* 181–199.
Kendler, K. (2005). A gene for: The nature of gene action in psychiatric disorders. *American Journal of Psychiatry, 162,* 1243–1252.
List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy, 106*(9), 475–502.
Maslen, C. (2009). Proportionality and the metaphysics of causation. *Philsci Archive*. http://philsci-archive.pitt.edu/4852/. Accessed 7 Jan 2015.
Menzies, P., & List, C. (2010). The causal autonomy of the special sciences. In C. McDonald & G. McDonald (Eds.), *Emergence in mind*. Oxford: Oxford University Press.
Shapiro, L., & Sober, E. (2012). Against proportionality. *Analysis, 72,* 89–93.
Sloman, S., & Lagnado, D. (2005). Do we 'do'? *Cognitive Science, 29,* 5–39.
Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search*. Cambridge: MIT Press.
Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science, 71,* 833–845.
Weslake, B. (2010). Explanatory depth. *Philosophy of Science, 77,* 273–294.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. New York: Oxford University Press.
Woodward, J. (2008). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: New essays on reduction, explanation, and causation* (pp. 21—262). Oxford: Oxford University Press.
Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy, 25,* 287–318.
Woodward, J. (2015a). Interventionism and causal exclusion. *Philosophy and Phenomenological Research, 91,* 303–347.
Woodward, J. (2015b). Methodology, ontology, and interventionism. *Synthese, 192,* 3577–3599.
Woodward, J. (2016a). The problem of variable choice. *Synthese, 193,* 1047–1072.
Woodward, J. (2016b). Unificationism, explanatory internalism, and the autonomy of the special sciences. In J. Pfeifer & M. Couch (Eds.), *The philosophy of philip kitcher* (pp. 121–146). Oxford: Oxford University Press.
Woodward, J. (2017a). Intervening in the exclusion argument. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a difference: Essays on the philosophy of causation* (pp. 251–268). Oxford: Oxford University Press.
Woodward, J. (2017b). Explanation in neurobiology: An interventionist perspective. In D. Kaplan (Ed.), *Integrating psychology and neuroscience: Prospects and problems*. (pp. 70–100). Oxford: Oxford University Press.
Yablo, S. (1992). Mental causation. *Philosophical Review, 101,* 245–280.
Yablo, S. (1997). Wide causation. *Philosophical Perspectives, 11,* 251–281.