THIS IS A PRE-PRINT – PLEASE CITE THE VERSION OF RECORD AVAILABLE AT DOI: 10.1093/BJPS/AXY030

Explanatory Abstraction and the Goldilocks Problem: Interventionism Gets Things Just Right

Thomas Blanchard

British Journal for the Philosophy of Science, 71(2), June 2020

Abstract

Theories of explanation need to account for a puzzling feature of our explanatory practices: the fact that we prefer explanations that are relatively abstract but only moderately so. *Contra* Franklin-Hall ([2016]), I argue that the interventione1ist account of explanation provides a natural and elegant explanation of this fact. By striking the right balance between specificity and generality, moderately abstract explanations optimally subserve what interventionists regard as the goal of explanation, namely identifying possible interventions that would have changed the explanandum.

Suppose that a pigeon has been conditioned to peck when presented with red stimuli, as opposed to stimuli of other colours. On some occasion the pigeon is presented with a scarlet stimulus. Consider these two explanatory claims:

(1) The pigeon pecked because it was presented with a scarlet stimulus

(2) The pigeon pecked because it was presented with a red stimulus

Intuitively, (2) better explains the pecking, the reason being that it abstracts away from irrelevant details about the precise shade of red that the stimulus happened to have. While this case (due to Yablo [1992]) is only a toy example, it illustrates a well-known aspect of our explanatory practices: in many contexts, we tend to prefer relatively abstract explanations that filter out idiosyncratic details of the case at hand.[1] But an equally noteworthy feature of our explanatory practices is that

[1] One aspect of this phenomenon is that the explanations provided by `high-level' sciences such as biology or psychology are often deemed more illuminating than the explanations provided by lower-level sciences (for example, in many contexts we tend to prefer psychological to neurological explanations of intentional actions). It is

this preference extends to moderately abstract explanations only. To illustrate, imagine with Franklin-Hall ([2016]) that the pigeon would also have pecked if it had been tickled or offered food rather than presented with a red stimulus. The claim that

> (3) The pigeon pecked because it was either presented with a red stimulus, or tickled, or offered food

is intuitively a worse explanation than (2), the reason being that it is too abstract or unspecific: it doesn't say enough about the causal processes that led to the pecking to fulfil our explanatory needs. A theory of explanation should thus help us understand why we prefer explanations that are `just right': neither too detailed nor too abstract. Weatherson ([2012]) aptly dubs this puzzle the `goldilocks problem'.

In this paper I argue that the interventionist account of explanation developed by Woodward and others elegantly solves the puzzle. To make this case I will proceed by examining Franklin-Hall's ([2016]) recent argument to the effect that interventionism's prospects at solving the goldilocks problem are especially poor. Franklin-Hall's target is Woodward's ([2008], [2010]) `proportionality' account of explanatory abstraction, whose central idea is that relatively abstract explanations do a better job than comparatively more detailed explanations at describing the pattern of dependence of the explanandum on its causes. According to Franklin-Hall, this account dramatically overshoots and ends up counting overly abstract explanations as optimal. I will argue that Franklin-Hall's attack relies on a misconstrual of the interventionist proportionality desideratum; properly understood, the desideratum correctly prefers moderately abstract explanations to overly detailed ones without counting dizzyingly abstract explanations as optimal. Moreover, as I will further argue, interventionism can naturally explain why overly abstract explanations are unsatisfactory: they do a poor job at meeting what interventionists regard as the fundamental goal of explanation, namely identifying interventions that would have changed the explanandum. On an interventionist picture, moderately abstract explanations thus emerge as optimal because they are specific enough to precisely identify which factors could have been `wiggled' to change the explanandum, and abstract enough to fully capture the pattern dependence of the outcome on these factors.

A few preliminary remarks are in order. First, I take the defining feature of explanatory abstraction to be this: an explanation E1 is more abstract than another E2 when the explanans cited by E2 entails the explanans cited by E1 but not *vice versa*. Thus (2) is more abstract than (1) because the pigeon being presented with a scarlet stimulus entails that it was presented with a red stimulus, but not *vice versa*. ((2) in turn is less abstract than (3), as is easy to verify.) This definition makes precise the idea that a more abstract explanation contains fewer details about the circumstances responsible for the explanandum. Second, my examination of the goldilocks problem is limited to explanations of singular events that cite explanatory factors that are causally relevant to the explanandum (like (1), (2), and (3) do). So I will not explore the issue of explanatory

widely thought that such explanations are generally superior precisely because they abstract away from irrelevant details.

abstraction in the context of general explanation[2], and I will also leave aside the question whether and how it arises in cases involving non-causal explanations.

1 Interventionism, Proportionality, and Franklin-Hall's objection

Interventionists take causation to be a matter of difference-making, understood as counterfactual dependence under certain ideal manipulations or *interventions*: a cause makes a difference to its effect insofar as an intervention on the cause would be associated with a change in the effect. To make this idea precise, it is best to represent causal relata using variables. In the case of singular or `actual' causation (the only one of interest here), the values of the relevant variables represent the occurrence or non-occurrence of some singular event or circumstance. (While values of variables represent causal relata, it will often be convenient and harmless to drop talk of representation and speak directly of the value of a variable causing (or explaining) the value of another variable.) To illustrate, return to Yablo's example; let *Red* take value 1 if the stimulus is red and 0 if it is of some other colour, and let *Peck* take value 1 if the pigeon pecks and 0 otherwise. Then for interventionism the claim that the stimulus being red caused the pigeon to peck can be cashed out as the claim that *Peck* would have taken value 0 under an intervention setting *Red* at 0 (and doing nothing else[3]). How to properly define the notion of an intervention is a subtle matter (see Woodward [2003], ch. 3). For our purposes it is enough to think of an intervention as a causal process that sets a variable at a certain value in a way that renders it independent from its usual causes, and doesn't directly affect any other part of the relevant causal structure besides the target variable.

This account of actual causation, note, is only a first pass. It doesn't apply to well-known cases of pre-emption and overdetermination where the cause doesn't make a difference to its effect**.** I will briefly return to the question of how to extend the interventionist account to those cases in section 2, but for the most part the issues discussed later on in the paper can be addressed using the simplified account of causation just presented, as we will see.

Turning now to explanation, interventionists regard (causal) explanation as aiming to provide information about the explanandum's causes understood along the lines above. More specifically, causal explanation involves exhibiting how the explanandum's occurrence depended on its causes. Given how interventionists understand causal dependence, this involves identifying interventions on those causes under which the explanandum would have been different. In short, for interventionism the goal of causal explanation is to identify explanandum-changing interventions. As Woodward puts it:

[2] In the case of explanation of generic patterns there is arguably no preference for moderate over maximal abstraction. For instance, a response to the question "Why do pigeons peck?" that mentions all possible ways in which peckings may occur ("Because they have been habituated to peck at certain stimuli, or tickled, or...") strikes me as a perfectly appropriate explanation. See Woodward ([unpublished]) for similar remarks.

[3] This qualification is important. Although an intervention that changed both the colour of the stimulus *and* the colour of the Golden Gate Bridge to non-red would have changed the value of *Peck*, this doesn't mean that the colour of the bridge made a difference to the pigeon pecking. A feature of the world is a difference-maker for an outcome only if intervening on this feature and just that feature would be sufficient to change the outcome.

> We explain an outcome by identifying conditions under which the explanandum-outcome would have been different… [S]uccessful causal explanation consists in the exhibition of patterns of dependency… between the factors cited in the explanans and the explanandum – factors that are such that changes in them produced by interventions are systematically associated with changes in the explanandum. ([2008], p. 228)

More specifically, the interventionist theory of explanation has two components. The first is a set of conditions on minimally adequate explanations (see Woodward [2003], p. 203):

> **Min**: A minimally adequate explanation of an event $Y=y$ consists of
> 1. A true statement to the effect that some variable X distinct from Y actually took value x. x is X's `actual value', whereas other values of $X$ are its `contrast values'.
> 2. A causal claim of the form $Y=f(X)$ correctly describing the values Y would have taken under interventions on the various values of X, and which correctly entails that for at least some contrast value x' of X, Y would have taken some value $y' \neq y$ had an intervention set X at x'.

While explanations are usually not explicitly couched in the language of variables and equations, it is often fairly easy to reconstruct them in this framework. To illustrate consider the explanation of the pecking (2). A request for an explanation of the pecking is naturally interpreted contrastively, as asking why the pigeon pecked rather than not pecked. So the obvious choice of explanandum variable is the variable *Peck*. And (2) is naturally read as saying that *Peck* took value 1 instead of 0 because *Red* took value 1 rather than 0. Recast in the format of variables and equations, the content of (2) thus becomes

> (2*) *Red*=1 and *Peck*=*Red*

And (2*) satisfies **Min**, as is easy to verify.

Any explanation that satisfies **Min** identifies at least some interventions that by changing certain variables would have changed the explanandum, and thereby goes at least some way towards satisfying what interventionists see as the purpose of explanation. Yet as will become clear below, two minimally adequate explanations of the same outcome may differ in how well they fulfil that purpose. Thus the second component in the interventionist account of explanation is a set of further desiderata on good causal explanations.[4] One of them is the desideratum of proportionality. The intuition behind the notion of proportionality (originally due to Yablo (1992)) is that an explanation is better insofar as it invokes a causal variable that contains just enough detail to account for the effect, but no more. As Woodward spells it out, a (minimally adequate) explanation satisfies this desideratum when the following principle **P** is satisfied:

> **P:** (a) [the explanation] explicitly or implicitly conveys accurate information about the conditions under which alternative states of the effect will be realized and (b) it conveys

[4] See (Hitchcock and Woodward [2003]) and (Woodward [2010]) for a discussion of various such desiderata.

> 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. ([2010], p. 298)

Woodward ([2008], [2010]) further argues that the superiority of (moderately) abstract explanations over comparatively more detailed ones is explained by the fact that the former generally do a better job than the latter at meeting this desideratum.[5] To illustrate, consider the contrast between the two explanations of the pigeon pecking (1) and (2). (We assume that the content of (2) is captured in (2*) above.) In a number of places, Woodward ([2008], pp. 235–6; [2010], pp. 297–9) suggests that the default reading of (1) is as asserting that the presentation of a scarlet rather than non-scarlet stimulus caused the pecking. On this default interpretation, (1) involves a binary cause variable – call it $Scarlet_{Default}$ – whose value 1 represents the presentation of a scarlet target and 0 the presentation of a non-scarlet target, together with the causal claim $Peck=Scarlet_{Default}$. So understood, (1) is clearly inferior to (2*), as the former violates condition (a) in **P** while the latter doesn't: it implies falsely that changing the stimulus's colour to a non-scarlet but still red colour would have changed *Peck*'s value, thereby conveying inaccurate information about the conditions under which the explanandum would have been different.[6] True, other choices of `scarlet' variables are possible which – although they might not necessarily constitute natural reconstructions of (1) – do not have this defect. Consider for instance a many-valued variable $Scarlet_{Many}$ whose value 1 represents the presentation of a scarlet stimulus and whose other values each represent the presentation of a stimulus of a slightly different colour shade. Together with a causal claim mapping every value of $Scarlet_{Many}$ representing a shade of red onto *Peck*=1 and every other value to *Peck*=0, this yields an explanation which asserts only true counterfactuals about the value *Peck* would have taken under various interventions on the stimulus's colour. But **P** still entails that this explanation is inferior to (2*), because the former (and not the latter) violates condition (b) in **P**. While the $Scarlet_{Many}$ explanation explicitly makes distinctions between possible colours of the target that fail to be associated with changes in the outcome, (2*) distinguishes only between those states of the target associated with alternative outcomes, and thereby more elegantly and parsimoniously describes the dependence of the pecking on the target's possible colours (Woodward [2010], p. 298).

Now, as Franklin-Hall ([2016], p. 564) notes, there is one way to construct a scarlet explanation of the pecking that perfectly meets the requirements of **P**, namely by using a binary variable whose contrast value represents some alternative colour shade that is not associated with pecking, for instance a binary variable whose value 1 represents the stimulus being scarlet and whose value 0 represents the stimulus being cyan. Since most of what follows will be devoted to

[5] In the interventionist framework, an explanation E1 is more abstract than another E2 when E2's explanans variable taking its actual value entails that E1's explanans variable takes its actual value, but not *vice versa*.

[6] As a reviewer helpfully pointed out, it would be improper to claim (as I did in earlier versions of this paper) that the $Scarlet_{Default}$ explanation isn't even minimally adequate, on the ground that the causal claim $Peck=Scarlet_{Default}$ yields the wrong value for *Peck* under certain interventions on the cause variable. According to Woodward, an adequate explanation need not yield the right value for the effect variable under all interventions. (In Woodward's terminology, a minimally adequate explanation can have limited `invariance'.) The idea that a good explanation should not convey inaccurate information about the values that the effect variable would take under various interventions is a further desideratum over and above the requirements for minimal explanatory adequacy.

compare the relative merits of *Red* and this variable, I will simply call it `*Scarlet*'. This choice of variable allows us to formulate the following explanation of the pecking[7]:

(1*) *Scarlet*=1 and *Peck*=*Scarlet*

(1*) is minimally adequate, and satisfies the letter of **P**. Nevertheless, from an interventionist standpoint this explanation is naturally regarded as still inferior to (2*). By contrast to (2*), (1*) remains silent on the outcomes associated with interventions setting the stimulus at possible colour shades other than scarlet and cyan, and thus doesn't reveal the full pattern of counterfactual dependence of the pecking on the stimulus's colour. That is, (1*) does a poorer job than (2*) at meeting what one may call an `exhaustivity desideratum' on good explanations - a desideratum to which Woodward alludes when he writes that (1) `tells us less than we would like to know about the full range of conditions under which the pigeon will peck or not' ([2008], p. 236). Although this desideratum isn't explicitly encapsulated in **P**, it is clearly part of the spirit of the proportionality desideratum.

To sum up, the interventionist proportionality account of the superiority of abstract explanations is that overly detailed explanations either inaccurately describe the pattern of dependence of the explanandum on the cause, or describe this pattern in a less elegant or exhaustive way than a comparatively more abstract explanation.[8] Yet according to Franklin-Hall, this account faces a crippling objection: it systematically overshoots and entails that overly abstract explanations are superior to the moderately abstract ones favoured by our explanatory practices. The culprit is the exhaustivity desideratum. Taking her cue from Woodward's remark cited in the previous paragraph, Franklin-Hall contends that exhaustivity `requires that the cause variable's values collectively exhaust the causal possibility space, that is, the range of circumstances by which the explanandum event—as well as its contrast—might be brought

[7] One way to communicate this explanation in ordinary language is to use explicitly contrastive wording: `The pigeon pecked because the stimulus was scarlet rather than cyan'.

[8] Woodward ([unpublished]) proposes a slightly revised definition of proportionality. On this new formulation **P***, proportionality favours explanations involving explanatory variables that `more fully represent or exhibit those patterns of dependence that hold with respect to [the explanandum]' ([unpublished], sect. 3). `More fully' means in part `more exhaustively', so that **P*** makes explicit the desideratum of exhaustivity left implicit in **P**. A noteworthy feature of **P*** is that it doesn't include condition (b) of **P**, which favours explanatory variables that map 1-to-1 onto the effect variable. Here Woodward follows Shapiro and Sober ([2012]), who argue that this condition is implausibly strong. Their example involves two real-valued variables *X* and *Y* such that both *X*=3 and *X*=22 both map onto *Y*=6. Here it seems implausible to say that the claim `*Y* took value 6 because *X* took value 3' is inferior to the claim `*Y* took value 6 because *X* took either value 3 or value 22'. By abandoning condition (b) of **P**, **P*** doesn't imply anymore that an explanation of the pecking in terms of *Red* is preferable to an explanation in terms of $Scarlet_{Many}$. Indeed, on Woodward's new view, there is no deep reason to regard the former as better than the latter – although, as Woodward also points out, the reverse is also true, so that **P*** still licenses us to use the latter. I'm not entirely convinced that Shapiro and Sober's example should lead us to abandon the claim that the *Red* explanation is better than the $Scarlet_{Many}$, but I shall not try to argue this point here as it is largely orthogonal to the main issue of the paper. Even understood along the lines of **P*** rather than **P**, proportionality still helps us make sense of our preference for (2) over (1), as it rules that the various ways to construct a `scarlet' variable yield explanations that are either no better than or straightforwardly inferior to (2).

about.' ([2016], 566). But now consider a variable *Full* taking value 1 if the pigeon is either presented with a red stimulus, or tickled, or offered food, and 0 otherwise. (*Full* is so-called because – I'll stipulate – its value 1 represents all the possible circumstances conducive to pecking.) Explanation (3) can then be recast in interventionist terms as follows:

(3*) *Full*=1 and *Peck*=*Full*

(3*) satisfies **Min** and **P.** Moreover, it fares much better than (2*) with respect to exhaustivity: whereas (2*) describes only one circumstance conducive to pecking, (3*) describes all such circumstances, and is thus maximally exhaustive. The interventionist proportionality standard thus entails not only that (2*) is superior to (1*), but also that (3*) is superior to (2*). As Franklin-Hall notes, the argument easily generalizes. Since virtually all events can be brought by many different possible circumstances, proportionality entails that any such event is best explained via a dizzyingly abstract variable whose actual value represents the occurrence of at least one of these possible circumstances, and whose contrast value represents their joint non-occurrence.

If so, then for interventionism to have any prospect at solving the goldilocks problem, the proportionality standard needs to be supplemented with a `downward-pulling' constraint on appropriate explanations that can counterbalance proportionality's tendency to overshoot, either by ruling out overly abstract explanations from the start or by trading off with proportionality in such a way that moderately abstract explanations emerge as optimal. However - and this is the second part of Franklin-Hall's argument – it is not at all clear what such a constraint might be. A natural thought is that there is something artificial or defective about (3*)'s explanatory variable: *Full* doesn't seem to carve up the causal structure of the situation in the right way, improperly mixing up circumstances of very different kinds.[9] The challenge is to spell out what `carving up reality in the right way' amounts to here exactly. One obvious suggestion is to appeal to Lewis's ([1983]) thesis that certain events are more metaphysically natural than others, the idea being that the disjunctive event represented by *Full*'s actual value isn't natural enough to figure in good explanations. Yet this suggestion faces well-known difficulties: for instance, it is unclear how we could ever get epistemic access to natural properties and what entitles us to regard current scientific theories as latching onto them (Loewer [1996]). Moreover, interventionists explicitly endorse a very liberal view about the sorts of events or circumstances that can enter into causal (and hence explanatory) relations. Within the interventionist framework, the only constraint on causal relata is the familiar requirement that they be `distinct' in Lewis's ([1986b]) sense; that is, logically and conceptually independent of each other. That is, every value of the cause variable should be logically and conceptually compossible with every possible value of the effect variable.[10] Besides this constraint, interventionists make no further demand on the sorts of events that can be causal relata. Woodward, in particular, is explicit that the theory does not include any

[9] Thanks to a reviewer here for helping me articulate this point.

[10] See for example (Halpern and Hitchcock [2010], p. 397) and (Woodward [2015], p. 310). This requirement is needed because without it, interventionism (or any counterfactual theory of causation for that matter) is vulnerable to an obvious objection, namely that it makes (say) John crying a cause of him crying loudly on the ground that the latter depends on the former, despite the fact that the relationship here is conceptual rather than causal.

substantive metaphysical commitment about the ontological status of causal relata.[11] A heavy-duty `naturalness' constraint on causal relata would thus be at odds with the metaphysically lightweight character of interventionism – an aspect of the view that many find very attractive.

[11] See for instance (Woodward [2016]).

2 Exhaustivity Reconsidered

Thus, if Franklin-Hall is right, interventionism's prospects of offering an adequate solution to the goldilocks problem are particularly poor. Yet further examination yields the opposite verdict, or so I will now argue.[12]

My first step will be to examine the interventionist exhaustivity desideratum in more detail, and to show that interventionists need not and should not endorse Franklin-Hall's formulation of this desideratum. To see why, we need to consider what exactly motivates a preference for more exhaustive explanations from an interventionist standpoint. In the interventionist literature to-date one finds very little explicit discussion of what this desideratum amounts to and the motivations for it, besides Woodward's passing remark that (1) `tells us less than we would like to know about the full range of conditions under which the pigeon will peck or not' ([2008], 236). But while this remark may suggest that interventionism is indeed committed to Franklin-Hall's version of exhaustivity, this is in fact not the case.

In order to show this, let's return to the case of (2*) vs. (1*) that motivated the introduction of the exhaustivity desideratum in section 1, and let's examine why interventionists should regard (1*) as inferior to the more exhaustive (2*). The answer, I suggest, goes like this. Remember that for interventionism, the goal of explanation is to identify explanandum-changing interventions – interventions that by changing certain aspects of the world would also have changed the explanandum. Thus if (1*) is explanatory, it is because it correctly identifies such a kind of intervention – namely interventions setting the stimulus's colour to cyan. But note that any intervention changing the stimulus's colour to cyan is necessarily an intervention changing it to a non-red colour. And (2*) tells us that any intervention of the latter kind would have changed the explanandum. So (2*) captures the fact that changing the stimulus's colour to cyan would have prevented the pecking, and hence contains all the explanatory information encapsulated in (1*). But (2*) also contains additional explanatory information that (1*) leaves out. First, (2*) captures a wider range of explanandum-changing interventions than (1*). Specifically, (2*) says that *any* intervention making the stimulus non-red would have prevented the pecking, whereas (1*) simply remains silent on what would have happened under changes to non-red shades other than cyan. Second, by contrast to (1*), (2*) explicitly tells us that the pecking would still have occurred had the stimulus been set to some shade of red other than scarlet, and hence that only changes to a non-red colour would have changed the outcome. It thereby does a better job than (1*) at singling out the exact range of changes to the stimulus's colour under which the outcome would have been different, and at separating them from interventions on the stimulus's colour that would leave the outcome unaffected.

Thus there is an obvious interventionist rationale for preferring (2*) to (1*), namely that the former is strictly more explanatory than the latter. That is, (2*) contains all the explanatory information included in (1*), while also containing additional explanatory information not contained in (1*). The superiority of the more exhaustive (2*) is thus a straightforward consequence of the fact that when it comes to explanatory information, the more the better.

[12] Woodward himself has recently offered a response to Franklin-Hall (Woodward [unpublished]), to which I'll return briefly at the end of the paper. The response I will offer on behalf of the interventionist is rather different than Woodward, although in my view the two are compatible and in fact nicely complement each other, as we'll see.

However, and crucially for our purposes, this rationale for preferring more exhaustive explanations does not apply in the case of (3*) vs. (2*): here we do not find that (3*) is strictly more explanatory than (2*), in the sense of containing all the explanatory information contained in (2*) and more besides. True, (3*) does contain explanatory information not included in (2*), insofar as it identifies certain explanandum-changing interventions not captured by (2*). Specifically, (3*) tells us (correctly) that the pigeon wouldn't have pecked under any intervention setting *Full* at 0 – that is, any intervention ensuring not only that the pigeon was not presented with a red stimulus, but also neither tickled nor fed. (2*), on the other hand, tells us nothing about the effects that such interventions would have on *Peck*. But (2*) also contains explanatory information that (3*) leaves out. Specifically, one piece of explanatory information contained in (2*) but not in (3*) is that merely changing the stimulus's colour to non-red and doing nothing else would have prevented the pecking. (3*) tells us nothing about the effects that such an intervention would have had, as the only explanandum-changing interventions this explanation identifies must do more than just setting *Red* at 0 – they must also ensure that the pigeon is neither tickled nor fed. Another way to put the point is that for all that (3*) tells us, the pecking might have been produced by an episode of tickling or feeding, in which case a mere intervention setting *Red* at 0 wouldn't have changed *Peck*'s value. So (3*) remains entirely silent about the effects that a mere intervention setting *Red* at 0 would have had: for all that it says, such an intervention may or may not have prevented the pecking. Here there is a key difference with the case of (2*) vs. (1*). Whereas (2*)'s explanans variable can be used to represent the kind of explanandum-changing interventions picked out by (1*) (as an intervention setting *Scarlet* at 0 is necessarily an intervention setting *Red* at 0), (3*)'s explanans variable is unable to represent the kind of explanandum-changing interventions identified by (2*), as intervening to set *Red* at 0 and doing nothing else is not a way to intervene to setting *Full* at 0, but a different kind of intervention altogether.

Two remarks are in order here to forestall possible misunderstandings. First, note that since in the actual situation in which the pecking took place the pigeon was neither tickled nor fed, it follows that *Full* would have taken value 0 under an intervention setting *Red* at 0 and doing nothing else. Isn't just setting *Red* at 0 just a way to set *Full* at 0? The answer is that although *Full* would have taken value 0 under an intervention setting *Red* at 0 and doing nothing else, this doesn't mean that the latter is an intervention *on* Full. An intervention on a variable X, remember, must set X at a certain value in a way that overrides the normal causal structure and renders X independent of its other causes. This means that a causal process counts as an intervention on *Full* only if whether the pigeon is tickled or fed is entirely controlled by this causal process. But an intervention setting *Red* at 0 and doing nothing else doesn't causally affect whether the pigeon is tickled or fed. Thus, even though in the actual situation an intervention setting *Red* at 0 would have been sufficient to set *Full* at 0, this isn't enough to qualify the intervention as an intervention on *Full*.

Second, to fully appreciate the point under consideration, it is worth explicitly distinguishing the interventionist view of explanation from a superficially similar view on which the goal of explanation is to identify conditions sufficient for the explanandum not to occur. On *that* view of explanation, (3*) is clearly more explanatory than (2*). After all, by itself *Red* taking value 0 is not sufficient for *Peck* to take value 0; it is only if the pigeon is also neither tickled nor fed (that is, if *Full* takes value 0) that the pecking is guaranteed not to occur. But for

interventionism, explanation doesn't aim at identifying conditions sufficient for the explanandum's non-occurrence; rather, it aims at identifying hypothetical *interventions* under which the outcome wouldn't have occurred. And those two are not the same thing, as our example illustrates: although *Red* taking value 0 isn't by itself sufficient for *Peck* to take value 0, in the actual situation where the pecking took place a mere intervention setting *Red* at 0 and doing nothing else would have sufficed to prevent the pecking, since other conditions required for the pecking (namely that no tickling or feeding occurs) were already in place. Because (2*) captures this fact while (3*) doesn't, on interventionism there is no compelling reason to regard the latter as strictly more explanatory than the former.

The upshot of these considerations is that from an interventionist point of view, there is an important dissymmetry between our two cases. In the case of (1*) vs. (2*), there is a compelling reason for interventionists to regard the more exhaustive explanation (2*) as explanatorily superior, namely that it contains all the explanatorily relevant information contained in (1*) and more besides. But there is no similar compelling reason to prefer (3*) over (2*), as despite being the less exhaustive of the two the latter explanation nevertheless contains explanatory information left out of the former. This means that interventionists can coherently hold that (2*) is superior to (1*) without being forced to regard (3*) as even better.

Here an important worry needs to be addressed.[13] I have so far proceeded on the assumption that (2*) faithfully represents the content of (2). This is to assume that by telling us that the redness of the stimulus was the singular or `actual' cause of the pecking, (2) thereby tells us that changing the stimulus's colour to non-red would have prevented the pecking. But as mentioned earlier, the underlying interventionist definition of actual causation is a simplified one. That an event $X=x$ is an actual cause of another $Y=y$ does *not* in fact entail that a mere intervention on X would necessarily have changed Y's value, as cases of pre-emption and overdetermination reveal. Thus, on any refined interventionist definition of actual causation that can handle such cases, it will turn out to be the case that (2) does *not* entail that a mere intervention on *Red* would have prevented the pecking. Hence the piece of explanatory information contained in (2*) but omitted in (3*) isn't contained in (2) itself! One may therefore worry that while the argument I have just put forward shows that an interventionist may coherently regard (2*) as superior to (1*) without also counting (3*) as optimal, this does nothing to illuminate our pattern of preferences in the case of the ordinary language explanations (1), (2) and (3) introduced at the outset of this paper.

This is a fair worry, but one that can be answered. The key is to note that a refined interventionist account of causation that can handle cases of pre-emption and overdetermination will still entail that (2) does contain a piece of explanatory information omitted in (3), although one of a slightly different sort. Let me explain why. The project of constructing such a refined account is a complex and ongoing one, and there is as of yet no consensus on what it should look like exactly. (See Hitchcock [2001], Woodward [2003]; Halpern and Pearl [2005]; Hitchcock [2007]; Halpern and Hitchcock [2010], [2015]). But there is a consensus on the general form that it should take. To illustrate, consider the following case of overdetermination.[14] A measure will pass (M=1) as long as at least one of two people and vote in its favour. As a matter of fact, both vote in favour

[13] Thanks to a reviewer for pressing me to address this worry.

[14] Which I borrow from Halpern and Pearl ([2005]).

of the measure ($Vote_1$=1 and $Vote_2$=1). All the aforementioned proposals agree that $Vote_1$ is a cause of M=1 because *if* $Vote_2$ had taken value 0, an intervention setting $Vote_1$ at 0 would have set M at 0. More generally, virtually all these proposals agree that what makes $X=x$ is an actual cause of $Y=y$ is the existence of (possibly non-actual) circumstances or `contingencies' in which Y would have taken a different value under an intervention setting X at a different value (and doing nothing else), at least if the relevant contingencies satisfy certain conditions. What these proposals disagree about are the nature of the conditions that such ``causation-revealing' contingencies must satisfy.[15] We need not explore this question here. The key point for our purposes is that on a refined interventionist account of actual causation, we can expect the contents of (2) and (3) respectively to be synonymous with propositions of the following form:

> (2**) There exists a contingency satisfying certain conditions D such that had it obtained, *Peck* would have taken value 0 under an intervention setting *Red* at 0 (and doing nothing else).
>
> (3**) There exists a contingency satisfying certain conditions D such that had it obtained, *Peck* would have taken value 0 under an intervention setting *Full* at 0 (and doing nothing else).

And it is easy to see that (3**) doesn't entail (2**), so that (2**) does contain explanatory information omitted in (3**). The reason, once again, is that an intervention setting *Red* at 0 and doing nothing else is *not* an intervention setting *Full* at 0. So the claim that there exist some contingencies under which the latter kind of intervention would have changed the explanandum doesn't entail the existence of contingencies under which the former kind of intervention would have changed the explanandum. On the other hand, on a refined interventionist account of actual causation an explanation of the pecking in terms of *Scarlet*=1 will have the following content:

> (1**) There exist a certain contingency satisfying certain conditions D such that had it obtained, *Peck* would have taken value 0 under an intervention setting *Scarlet* at 0 (and doing nothing else).

But (2**) *does* also contain that information, as any contingency in which changing the stimulus's colour to non-red would have changed *Peck*'s value is also a contingency in which changing the stimulus's colour to cyan would have changed *Peck*'s value. (2**, of course, also contains additional explanatory information, for example that in the relevant contingency changing the stimulus's colour to cyan or any other non-red shade would have prevented the pecking.) The previous conclusion still holds: (2**) contains all the explanatory information contained in (1**) and more besides, but contains information omitted from (3**), so that the interventionist can

[15] To give an example of what such conditions may look like: Hitchcock's `WA' proposal ([2001], p. 290) requires these contingencies to be such that had they obtained, all the variables on the path between the cause and the effect would still have had the same value. In other words, a possible contingency is `causation-revealing' only if in that contingency the effect depends on the cause *and* the occurrence of the contingency would not have `perturbed' the causal path by which the cause leads to the effect.

coherently regard (2**) as better than (1**) without being forced to also regard (3**) as even better.[16] In other words, the reasoning outlined a few pages before will still hold (although in a slightly different form) even in the context of a fully adequate interventionist account of causation, whatever exact form such an account will take. For convenience in the paper, in the remainder of this paper I will stick with the simplified account of actual causation presented in the last section, and with the related assumption that the problem of explaining our pattern of preference in the case of (1), (2) and (3) can be tackled by examining the relations between (1*), (2*) and (3*).

It is helpful for our purposes to examine in more detail why exactly interventionists can regard the more exhaustive (2*) as superior to (1*) without being forced to also recognize (3*) as even better. The reason is this. True, both (2*) and (3*) are more exhaustive than (respectively) (1*) and (2*) in Franklin-Hall's sense: both identify more possible circumstances causally relevant to the explanandum and its contrast. Thus, (2*) tells us that the pecking may have been brought about either by the stimulus being scarlet (as (1*) already tells us), or by it being of some other shade of red; and (2*) also tells us that *Peck* would have taken value 0 not only if the stimulus had been cyan (as (1*) says) but also if it had been of any other non-red shade of colour. Likewise, (3*) tells us that the pecking might have been brought either by the presentation of a red stimulus (as 2*) already tells us), but also by a tickling or a feeding. And it also tells us that for *Peck* to take value 0 not only must the pigeon not be presented by a red stimulus, it must also be neither tickled nor fed. Nevertheless there is a crucial distinction in the type of additional circumstances that each explanation captures. In the case of (2*) vs. (3*), the additional circumstances captured by (3*) are *distinct* from the circumstances already captured by (2*), in the sense of `distinct' introduced in section 1: whether or not the pigeon is tickled or fed is logically and conceptually independent of whether it is presented with a red stimulus. That is, whether the pigeon is tickled or fed has no logical/conceptual impact on what value *Red* takes. In the case of (1*) vs. (2*), however, the additional possible circumstances causally relevant to *Peck* that (2*) identifies are *not* distinct from those that (1*) already captures. Specifically, those additional circumstances are all conceptually incompatible with (or as I will also put it, *alternatives* to) the circumstances represented by *Scarlet*'s values: the stimulus being of some non-scarlet shade of red is incompatible with its being either scarlet or cyan, and the stimulus being of some non-red shade other than cyan is also incompatible with its being either scarlet or cyan. Thus (3*) and (2*) display different forms of exhaustivity relative to (2*) and (1*) respectively: while the additional possible circumstances causally relevant to *Peck* captured by (3*) are distinct from those already captured

[16] This is also true on Weslake's ([forthcoming]) interventionist definition of actual causation, although here matters a bit more complicated. Like other proposals, Weslake's definition requires a cause to make a difference to its effect in at least some contingencies satisfying certain conditions, but also asks of a cause that it meet other requirements as well (called STRAND and DIF in his definition). So on his definition, (1), (2) and (3) imply certain statements of form (1**), (2**) and (3**) respectively, but also have other implications as well, namely that (respectively) *Scarlet*=1, *Red*=1 and *Full*=1 satisfy STRAND and DIF. But, as the reader may verify: (a) on Weslake's theory, the additional information contained in (3) besides (3**) doesn't entail (2**) either, so that there is still explanatory information contained in (2) but not in (3); (b) the claim that *Red*=1 satisfies STRAND and DIF entails that *Scarlet*=1 satisfies those conditions as well, so that whether or not one regards the claim that *Scarlet*=1 satisfies those conditions as *explanatory* information, (2) also contains that information (in addition to also containing the information encapsulated in (1**)).

by (2*), but the additional possible circumstances captured by (2*) are not distinct from but instead alternatives to those already captured by (1*).

The difference can also be expressed by introducing variables representing the additional circumstances captured by (3*) and (2*) and left out by (2*) and (1*) respectively. Thus let *Tickle* take value 1 if the pigeon is tickled and 0 otherwise, and *Food* take value 1 if the pigeon is offered food and 0 otherwise. *Full*, note, can be recast as a function of *Red* and these two variables: specifically, *Full* takes value 1 if either *Red, Tickle* or *Food* takes value 1, and 0 if all of them take value 1. Likewise, let *Red** take value 1 if the stimulus is of some non-scarlet shade of red and 0 if it is of a non-cyan, non-red shade of colour. *Red* can then be recast as a function of *Scarlet* and *Red**: specifically, *Red* takes value 1 if either *Scarlet* or *Red** takes value 1, and 0 if either *Scarlet* or *Red** takes value 0. The aforementioned difference between the way in which (3*) is more exhaustive than (2*) and the way in which (2*) is more exhaustive than (1*) is reflected in the fact that while *Tickle* and *Food* are distinct from *Red* (in the sense that every combination of values of those variables is logically and conceptually compatible with every value of *Red*), *Red** and *Scarlet** are not distinct: specifically, every value of *Scarlet* is conceptually incompatible with every value of *Red** and *vice versa*.[17]

It is this difference that explains why (2*) contains all the explanatory content included in (1*) while (3*) leaves out some explanatory information contained in (2*), and hence why interventionists can coherently regard (1*) as superior to (2*) without being forced to recognize (3*) as even better. On the one hand, (3*) says that *Peck* would have taken value 0 under an intervention setting *Full* at 0 – that is, an intervention ensuring that neither *Red*, *Tickle* or *Food* take value 1. But since *Tickle* and *Food* are distinct from *Red*, it is possible for *Red* to take value 0 while either *Tickle* or *Food* take value 1, so that merely setting *Red* at value 0 and doing nothing else doesn't by itself guarantee that *Full* take value 0. This is why (3*) doesn't capture the fact that intervening to set *Red* at 0 and doing nothing else would have prevented the pecking, and hence omits explanatory information contained in (2*). By contrast, because the stimulus being cyan is incompatible not only with it being scarlet but also with it being any shade of red whatsoever, by telling us that the pecking occurred because the stimulus was red (2*) captures the fact that setting it at cyan would have changed *Peck*'s value.[18] In that way, it is able to preserve the explanatory

[17] It should be noted that this point is independent of how exactly we individuate the additional circumstances captured by the relevant explanation. Here I have introduced two different variables (*Food* and *Tickle*) to represent the possible circumstances causally relevant to *Peck* represented by (3*) and not (2*). Alternatively, those circumstances could be represented with a single variable *Ticklefood* taking value 1 the pigeon is either tickled or fed and 0 otherwise. (*Food* then takes value 1 if either *Red* or *Ticklefood* takes value 1, and 0 otherwise.) The relevant point would still hold, as *Ticklefood* is distinct from *Red*. Likewise, instead of representing the circumstances potentially relevant to the pecking captured by (2*) but not (1*) with a single variable *Red**, we could instead introduce a range of binary variables, each of which has a value 1 representing the stimulus being of a certain shade of red and a value 0 representing the stimulus being of a certain shade of non-red, so that every possible colour other than scarlet and cyan can be represented via one of these variables. (*Red* can then be recast as a function of *Scarlet* and these variables: specifically, *Red* takes value 1 just in case either *Scarlet* or one of these variables takes value 1, and 0 if either *Scarlet* or one of these variables takes value 0.) Here again the point still stands, namely that these additional variables are not distinct from *Scarlet*.

[18] To put it in terms of variables: because every value of *Scarlet* is incompatible with every value of *Red**, a mere intervention setting *Scarlet* at 0 is enough to ensure not only that *Scarlet* doesn't take value 1, but also that *Red** doesn't take value 1, because *Scarlet*=0 and *Red**=1 are conceptually incompatible. Hence, in telling us that the

information showcased in (1*). At the same time, because (2*) also describes the value *Peck* would have taken under other possible colours alternative to scarlet and cyan, it communicates more information explanatorily relevant to the pecking. Specifically, by telling us that *Peck* would have taken value 0 had the stimulus been of any non-red colour, it captures the fact that a wider range of interventions that would have prevented the explanandum. And by telling us that *Peck* would still have taken value 1 had the stimulus been of any red shade, it captures the fact that *only* changes to a non-red colour would have prevented the pecking, and hence better singles out the exact range of changes to the stimulus's colour that would have changed the explanandum.

The upshot is that while Franklin-Hall's formulation of the exhaustivity desideratum is insensitive to the distinction just drawn between these two forms of exhaustivity (effectively counting both as explanatory valuable), it is only the second one exhaustivity in *alternative* circumstances – that has compelling explanatory value from an interventionist standpoint. I propose, then, that the interventionist exhaustivity desideratum be formulated as follows:

> **Exhaustivity**: Consider two minimally adequate explanations E1 and E2 of an outcome $Y=y$ with explanans variables X_1 and X_2 respectively. Suppose that E2 correctly tells us which value Y would have taken under interventions setting X_1 at its possible values, while also correctly describing which value Y would have taken under circumstances alternative to (that is, mutually exclusive with) those represented by X_1's values, had those alternative circumstances been brought about by interventions. Then ceteris paribus E2 is a better explanation than E1.

And adding this requirement to **P**[19] yields a version of the proportionality desideratum that favours (2*) over (1*), as the interventionist wishes, and without overshooting: because proportionality so-understood doesn't count the kind of increase in exhaustivity displayed by (3*) in comparison to (2*) as explanatorily valuable, the desideratum remains simply silent on which of these two explanations – if any – is the better one.

It is worth considering another example to further illustrate how proportionality so-understood works. Suppose we want to know why a window broke (*Window*=1) rather than remained intact (*Window*=0). And suppose that a rock was thrown at the window by Suzy with a velocity of 3.8m/s, and that in the relevant circumstances throwing the rock at any velocity higher than 3m/s would have been enough to break the window. Intuitively

(4) The window broke because Suzy threw the rock at a velocity of 3.8m/s

is too detailed an explanation of the window breaking: the rock's exact velocity is an irrelevant detail. The following more abstract explanation, which filters out this irrelevant detail, is a more satisfactory account of the breaking:

---

pecking occurred because *Red* took value 1 – that is, because either *Scarlet* or *Red** took value 1 -, (2*) still manages to capture the fact that the pecking wouldn't have occurred under an intervention setting *Scarlet* at 0.

[19] Or using this formulation to fill out what `fully' means in Woodward's new formulation of exhaustivity **P*** (see Footnote 8).

(5) The window broke because Suzy threw the rock at a velocity higher than 3m/s

And this is also the verdict delivered by the proportionality desideratum proposed here. (5) can be recast as telling us that *Window*=*Rock* and that *Rock*=1, where *Rock* is a variable taking value 1 just in case Suzy throws the rock at a velocity higher than 3m/s and 0 otherwise. On the other hand, the natural way to reconstruct (4) is as involving a binary variable whose actual value represents the rock having a velocity of 3.8m/s, and whose contrast value represents one or more possible alternative velocities of the rock lower than 3m/s.[20] For definiteness, let's assume that the contrast value in question represents the rock being thrown at 2m/s. Then the exhaustivity desideratum rules that (5) is superior to an explanation in terms of such a variable (or any similar one). The reason is that (5) correctly describes the value that *Window* would have taken under the full range of alternative possible velocities of the rock, including alternatives over which the less abstract explanation under consideration remains silent (namely any velocity other than 3.8 and 2m/s).

By contrast, compare (5) with an even less – and insufficiently – specific explanation of the window breaking such as

(6) The window broke because either Suzy threw a rock at it with velocity higher than 3m/s, or because a bomb exploded near the window

Franklin-Hall's version of the proportionality desideratum would have us regard (6) as an even better explanation than (5). By contrast, the proportionality desideratum as I propose to understand it doesn't overshoot in this way: it remains neutral on which of (5) or (6) is the best explanation. In particular**, Exhaustivity** gives us no reason to prefer (6) to (5). The reason is that the additional circumstances causally relevant to the breaking are not alternatives to those already cited in (5), but distinct from them: whether a rock was thrown at the window with a certain speed and whether a bomb exploded are logically/conceptually independent aspects of the world. And **Exhaustivity** simply doesn't say whether that form of exhaustivity is explanatorily valuable.

To put the results of this section in perspective, it is instructive to compare the interventionist proportionality desideratum with other popular principles that have been offered to explain the superiority of relatively abstract explanations, and which *do* overshoot. Thus consider the idea that more abstract explanations are better because they are more general: they can be applied to a wider range of (possible or actual) cases.[21] While this hypothesis explains our preference for (2) over (1) (as the latter can be applied to more instances of pecking) it also favours (3) over (2), since (3) can be applied to more cases still – namely peckings produced by ticklings

[20] If the contrast value also represents velocities higher than 3m/s, the explanation will yield a partially inaccurate description of the dependence of the window breaking on the rock's velocity. One may also use a many-valued variable whose contrast values each represent a specific velocity. In that case, condition (b) of **P** will entail that this explanation is still too detailed compared to (5).

[21] This line of thought appears in (Fodor [1974]) and (Garfinkel [1981]), among many others. See (Weslake [2010]) for a recent defence of this view.

or food offerings. Or consider the thought that the inferiority of inappropriately specific explanations is due to them citing events or properties that are in some way *unnecessary* for the outcome to happen.[22] Here again, this hypothesis yields a preference for (2) over (1). (Clearly the stimulus being scarlet is unnecessary for the pecking to occur: all that matters is that it be red). But it also entails that (3) is even better. After all, it wasn't really necessary for the explanandum's occurrence that the pigeon be presented with a red stimulus: by hypothesis, any event instantiating the disjunctive property cited in (3) would have sufficed to ensure the pecking. The reason these two principles overshoot is because they make exhaustivity as defined by Franklin-Hall an explanatory virtue: they entail that the more circumstances by which the explanandum might be brought about an explanation identifies, the better it is (either because it can be applied to more cases, or because it better highlights conditions that are causally/nomically indispensable for the outcome to occur.) For interventionism, by contrast, the quality of an explanation isn't directly tied to its ability to identify such circumstances, but to its ability to identify explanandum-changing interventions. And as the case of (3*) vs. (2*) illustrates, an explanation that fails to mention certain circumstances that may cause the explanandum may be able to represent certain explanandum-changing interventions that a more exhaustive explanation cannot represent. Identifying causally relevant circumstances and identifying explanandum-changing interventions associated with changes in the explanandum do not amount to the same thing, and it is the latter that interventionism cares about.

3 Interventionism and the Explanatory Value of Specificity

If the arguments of section 2 are correct, the proportionality desideratum explains why moderately abstract explanations are superior to inappropriately specific ones without overshooting. Yet while proportionality considerations do not favour overly abstract over moderately abstract explanations, they do not favour the latter kind of explanation over the former either. For instance, proportionality considerations by themselves do not entail that (2*) is superior to (3*); indeed, they do not allow us to rank one as better than the other. Thus the second aspect of the goldilocks problem – our preference for explanations that remain sufficiently specific about the actual circumstances responsible for the explanandum – remains to be addressed. But as I will now argue, interventionists have a good explanation of this phenomenon to offer, and one that doesn't presuppose any `naturalness' requirement or any other heavy-duty metaphysical assumption on the kinds of properties or events that can be cited in good explanations. Instead, the inferiority of insufficiently specific explanations can be traced back to the fact that such explanations do a less than satisfactory job at achieving what interventionists regard as the fundamental goal of explanation, namely identifying explanandum-changing interventions. Let me illustrate by returning to Yablo's pigeon example, focusing for now on the explanatory competition between (2*) and (3*).

In saying that (3*) does a less than satisfactory job at identifying explanandum-changing interventions, I am not denying that (3*) does manage to correctly pinpoint certain interventions under which *Peck* would have taken a different value. As we have seen, (3*) does correctly single out a certain range of interventions as explanandum-changing, namely interventions setting *Full*

[22] See (Strevens [2008]) for a development of this idea.

at 0. Instead, the problem with (3*), I suggest, is that by being highly unspecific about what actually happened, it suggests that some other interventions would have changed the explanandum, whereas they actually wouldn't have. Let me explain. Note that for all that (3*) tells us, the pigeon pecked not because it was presented with a red stimulus, but because it was tickled or fed instead. And if one of these two possibilities had been actual, then either intervening to prevent the pigeon from being tickled (and doing nothing else) or intervening to prevent it from being fed (and doing nothing else) would have been explanandum-changing. But of course, in the actual circumstances in which the pecking took place, none of these interventions would actually have changed *Peck*'s value. Thus, by telling us that the pecking might be due to a tickling or a feeding, (3*) suggests that some interventions may well have been explanandum-changing, whereas they actually were not. By contrast, no similar criticism can be voiced against the more specific (2*), which doesn't cite tickling or feeding as possible explanantia of the pecking. And on an interventionist view, this is a good reason to prefer the latter explanation: if the goal of explanation is to identify explanandum-changing interventions, it is better if an explanation doesn't actively suggest certain interventions as potentially explanandum-changing interventions if they actually were not.

It is helpful, I think, to rephrase this argument for the explanatory value of specificity using some new terminology. Say that a variable X is a `locus of explanandum-changing interventions' (for an explanandum $Y=y$) just in case intervening to change X's value - and doing nothing else – would have led Y to take a value other than y. The problem with (3*) can then be put like this. Since *Full*=1 is true just in case *Red*=1 ∨ *Tickle*=1 ∨ *Food*=1, by telling us that the pecking occurred because *Full* took value 1,(3*) leaves it open that the pecking occurred not because *Red* took value 1, but because either *Tickle* or *Food* did. And if this had been the case, then *Tickle* or *Food* would have been a locus of explanandum-changing interventions. But neither actually was: as is easy to verify, in the actual situation where the pecking took place any intervention changing *Tickle* or *Food* would have still resulted in *Peck* taking value 1. The aforementioned defect with (3*), then, is that by telling us that the pecking possibly occurred because either *Food* or *Tickle* took value 1, it suggests these variables as potential loci of explanandum-changing interventions for *Peck*=1, whereas they actually were not. (2*), by contrast, doesn't have this defect, since it doesn't cite *Tickle*=1 or *Food*=1 as possibly responsible for the pecking. And so, I claim, on interventionism it is no surprise that we regard (2*) as a better explanation of the pecking than (3*).

To put it in more general terms, the claim put forward here is that the interventionist theory of explanation naturally yields the following principle:

> **Specificity**: Consider two minimally adequate explanations E1 and E2 of an outcome $Y=y$. Suppose that E1 involves an explanans variable X_1 (with actual value $X_1=1$) and E2 an explanans variable X_2 (with actual value $X_2=1$). Suppose, moreover, that X_1 is a function of X_2 and other variables X_3, … X_n, such that X_1 takes value 1 just in case either $X_2=1$ or $X_3=1$ … or $X_n=1$. Finally, suppose that while X_2 was a locus of explanandum-

> changing interventions, X_3, … X_n were not. Then *ceteris paribus* E2 is a better explanation than E1.[23]

The rationale for that principle is that by being less specific than E2 and telling us only that the explanandum occurred because either X2 took value 1 or one of X_3, X_4, … or X_n did, E1 actively suggests these variables as potential loci of explanandum-changing interventions, whereas just intervening on any of them would have had no effect on the explanandum. And our preference for (2*) over (3*) is, I claim, an instance of this principle: since *Full* takes value 1 just in case *Red*=1 ∨ *Tickle*=1 ∨ *Food*=1, and since the latter two variables weren't loci of interventions that would have changed *Peck*'s value, **Specificity** favours (2*) over (3*).[24, 25] More generally, the hypothesis put forward here is that from an interventionist point of view, specificity is valuable to the extent that it helps us zoom in on actual loci of interventions that would change the explanandum, and that moderately abstract explanations strike us as better than more abstract competitors precisely because their higher specificity is explanatorily valuable in this way.[26]

This isn't to say that from an interventionist point of view any increase in specificity is *ipso facto* explanatorily beneficial. In particular, and despite what its name may suggest, **Specificity** doesn't indiscriminately favour more detailed explanations. For instance, it gives us no reason to regard (1*) as superior to (2*), despite the fact that the former gives us more

[23] This explanatory desideratum should not be confused with a quite different desideratum with the same name discussed by Woodward ([2010], pp. 301–14) which favours causal/explanatory variables such that, by varying the state of those variables, one can modulate the state of the explanandum in a fine-grained manner.

[24] Note that **Specificity** favours a more specific explanation E2 over a less specific E1 only if the actual value E1's explanans variable X_1 can be rewritten as a disjunction of the actual value of E2's explanans variable X_2 and values of *other* variables. In general there may be several ways to do so, depending on how one individuates the possible circumstances cited by E1 but not E2 as possible causes of the explanandum. But **Specificity**'s verdict is insensitive to which individuation scheme one chooses. For instance, **Specificity** still favours (2*) over (3*) if we choose to represent the additional circumstances relevant to *Peck* captured by (3*) but not (2*) with a single variable *Foodtickle* instead of two separate variables *Food* and *Tickle* (see Footnote 17). Since *Full*=1 is equivalent to *Red*=1 ∨ *Foodtickle*=1 and the latter variable wasn't a locus of explanandum-changing interventions, (2*) still comes out as superior to (3*) on this individuation scheme.

[25] Likewise, in the window-breaking example from section II, **Specificity** correctly favours (5) over the highly unspecific (6). To see this, let *Bomb* be a variable taking value 1 if a bomb explodes near the window, and 0 otherwise. (6) can then be represented as telling us that the window broke because either *Rock* or *Bomb* took value 1.And since *Bomb* wasn't a locus of explanandum-changing interventions for the window breaking, **Specificity** rules that (5) is a better explanation of the pecking.

[26] As in our discussion of exhaustivity in section II, taking into account considerations of overdetermination and pre-emption slightly complicates the picture, but not in a way that makes a substantial difference. The issue can be brought out by imagining a case where the pigeon is both presented with the red stimulus and tickled, so that the pecking is overdetermined. In such a case (2) is presumably still a better explanation than (3). Yet **Specificity** cannot explain why, as in that situation *Red* is not a locus of explanandum-changing interventions as I have defined that notion earlier. (Given that *Tickle* took value 1, merely intervening to set *Red* at 0 wouldn't have changed *Peck*'s value.) The obvious solution is to refine this definition by saying that a variable is a locus of explanandum-changing interventions when there exist `causation-revealing' contingencies (in the sense of the term introduced in section II) in which changing the variable's value would have changed the explanandum. In the present case, all interventionist accounts of actual causation count the contingency `*Tickle*=0' as causation-revealing for *Red*, and in that circumstance intervening to set *Red* at 0 would have changed *Peck*'s value. With this refined definition **Specificity** does count (2) as superior to (3) even in this case.

information about the actual circumstances that led to the pecking. True, just like *Full* can be represented as a disjunction of *Red*=1 and the values of other variables (*Tickle*=1 and *Food*=1), so can *Red*=1 be represented as a disjunction of *Scarlet*=1 and values of other variables. For instance, as we have seen in the previous section, *Red*=1 is equivalent to *Scarlet*=1 ∨ *Red**=1, where (remember) *Red** takes value 1 if the stimulus is of some shade of red other than scarlet and 0 if the stimulus is of some non-red shade other than cyan. But whereas *Tickle* and *Food* are not loci explanandum-changing interventions for the pecking, *Red* is* such a locus: as is easy to verify, any intervention setting *Red** at 0 would have changed *Peck*'s value to 0. So despite the fact that (1*) is a more specific explanation of the pecking than (2*), **Specificity** doesn't count the former as superior.[27]

Interestingly, the fact that **Specificity** doesn't favour (1*) over (2*) despite favouring (2*) over (3*) is due to the fact that, as we saw in section 2, these two cases differ in an important way: in the former but not the latter case, the additional possible circumstances causally relevant to the explanandum cited by the more abstract explanation are conceptually related to (specifically, mutually exclusive with) those cited by the less abstract explanation. This, remember, is reflected in the fact that the variable *Red** - which represents the additional circumstances cited by (2*) but not by (1*) – is not distinct from (1*)'s explanans variable *Scarlet*: every value of *Red** is incompatible with every value of *Scarlet* and *vice versa*. And this fact explains why it is perfectly appropriate for (2*) to tell us only that the pecking occurred because either *Scarlet* or *Red** took value 1. Because the values of those two variables are mutually incompatible, setting *any* of them at value 0 guarantees that neither takes value 1. Hence, as long as the pecking occurred because one of these variables took value 1, any intervention setting *any* of these two variables at value 0 is guaranteed to be explanandum-changing. Thus whether it was *Scarlet* or *Red** that took value 1, both variables are guaranteed to be loci of explanandum-changing interventions, and from an interventionist point of view there is no explanatory payoff in telling us which of these possibilities was actually the case.[28] Now, for the same reasoning to apply in the case of (2*) vs. (3*), it would have to be the case that setting anyone of the three variables *Red*, *Tickle* or *Food* at value 0 (and doing nothing else) is enough to ensure that neither takes value 1. (Then it would be perfectly appropriate to explain the pecking only by mentioning that one of these three variables took value 1, since either way all of them would be loci of explanandum-changing interventions.) But this isn't the case. And the reason is that *Red*, *Tickle* and *Food* are distinct from one another, so that intervening on just one of these variables puts no constraint on the values of the others. In particular, setting anyone of these variables at value 0 (and doing nothing else) isn't in itself sufficient to ensure that the others do not take value 1.

These considerations allow us to better understand which kind of specificity is explanatorily detrimental from an interventionist point of view. Specifically, what they suggest is that when an explanation E1 `abstracts away' from another explanation E2 by telling us that the outcome was caused either by the circumstance picked out by E2 or some other circumstances,

[27] As the reader may easily verify, the same result holds for other ways to represent the additional circumstances captured by (2*) and not (1*) (see Footnote 17).

[28] It is easy to see that in the window-breaking example from section 2 **Specificity** likewise doesn't favour (4) over the less specific explanation (5). And the reason, once again, is that the additional possible circumstances relevant to the explanandum captured by (5) are alternatives to those already captured by (4).

interventionism privileges the more specific explanation if – but only if - the latter circumstances are distinct from the one picked out as explanans by E2. On the other hand, if the circumstances under consideration are not distinct from but alternatives to the ones identified by E2, then the fact that E2 is more specific about which of these alternative circumstances was actually instantiated doesn't confer it an advantage over E1. Indeed, the results of section 2 show that in that case, the more abstract explanation is preferable, on the ground that it better satisfies the interventionist desideratum of proportionality.

These considerations also reveal that interventionism can do justice to the intuition that a variable such as *Full* fails to `carve reality in the right way', and without having to appeal to any heavy-duty metaphysical thesis about `natural' features of the world. On the view proposed here, *Full* is defective not because it picks out an `unnatural' feature of the world but because it mixes up aspects of the world – whether the stimulus was red, whether the pigeon was tickled and whether it was fed – that are distinct from one another in such a way that one can constitute a locus of explanandum-changing interventions without the others being one as well. And if the goal of explanation is precisely to identify explanandum-changing interventions, we have strong reasons not to collapse such distinct aspects of the world into a single variable when attempting to explain an outcome.

To complete this account of the value of specificity in the interventionist framework, we need to address an objection. The worry is that the interventionist requirement of specificity `undershoots', and ends up counting the moderately abstract explanations we regard as optimal as insufficiently specific. For instance, although **Specificity** doesn't favour (1*) over (2*), one may nevertheless worry that this requirement still yields the (counterintuitive) result that (2*) is a suboptimal explanation. To see this, note that there are many ways – all distinct from one another – in which the red stimulus presentation might have occurred: the pigeon might have been presented with a red tile, or with my red shirt, or with any other number of red object. (Let's assume it was the tile.) And by telling us only that the pigeon was presented with *a* red stimulus, (2) seems to leave it entirely open which of these circumstances actually occurred. This suggests that if (2*) is to be a faithful reconstruction of (2)'s content, we must interpret *Red*=1 in a way such that it is equivalent to a disjunction of the values of variables distinct from one another, only one of which was a locus for explanandum-changing interventions. For instance, let *Tile* take value 1 if the tile is red and 0 if it isn't, and *Shirt* take value 1 if my shirt is red and 0 if it isn't. Then *Red*=1 is equivalent to the disjunction of *Shirt*=1, *Tile*=1, and the values of many other variables. These variables are distinct from one another, and only one – *Tile* – was a locus of explanandum-changing interventions. Hence (2*) appears to have the defect that **Specificity** penalizes for – for instance, it suggests that intervening to change the colour of my shirt to non-red (and thereby setting *Shirt* at 0) would have changed the explanandum, whereas it actually wouldn't have. And yet the fact that neither (2) nor its natural reconstruction in the format of variables do not precisely identify which exact object the pigeon was presented with doesn't generally strike us as a particularly objectionable form of unspecificity for explanatory purposes.

However, this objection relies on a mistake, I think. It erroneously interprets the phrase `a red stimulus' in (2) as an existential generalization of the form `some red stimulus or other'. A more natural interpretation of this phrase is that it is used referentially, to designate the particular red object with which the pigeon was presented. Witness the fact that `The pigeon was presented

with a certain red stimulus' seems a perfectly appropriate paraphrase of (2). That is, (2) should not be read as saying that the pigeon was presented with some red stimulus or other, but as saying, *of* the particular stimulus with which the pigeon was presented, that it was red (thereby indicating that intervening on that very stimulus in order to make it non-red would have prevented the pecking).[29] If this is correct, it is therefore misleading to interpret *Red*=1 as equivalent to the disjunction *Tile*=1 ∨ *Shirt*=1 ∨… Instead, *Red*=1 is more appropriately read as rigidly referring to the particular stimulus that the pigeon was presented with and telling us that it was red, and hence as equivalent to *Tile*=1 (assuming, once again, that it was the tile that the pigeon was presented with).[30] So **Specificity** doesn't penalize (2*) after all.

Now, even if this is correct, it remains the case that, in some sense, (2) doesn't precisely identify which object the particular stimulus under focus was exactly. But note that in many contexts in which an explanation such as (2) might not be uttered (including the one in which I introduced this explanation as example at the beginning of the paper), the hearer isn't antecedently acquainted with the object that the pigeon was presented with. And so in such a context, there is simply *no* way for the speaker to identify that object other than by saying `a red stimulus'. It is therefore not surprising that we do not regard this lack of specificity as a defect, at least generally. As a case in point, consider a situation in which the audience to which the explanation is offered *does* have antecedent direct knowledge of the object that the pigeon was presented with. Imagine for instance that a member of the lab to which the pigeon belongs is explaining to another member why the pigeon pecked on a certain occasion, and suppose that both members know that the room in which the pigeon is stored contains two objects used as stimuli – a red triangle and a red square, say. And finally suppose that it was the triangle that the pigeon was presented. In that context, it would be rather unsatisfactory for the explainer to say that the pigeon was presented with a red stimulus (and nothing more), and preferable for her to say that it was presented with the red triangle. The reason is that since the audience is already acquainted with the triangle, putting things in this way better allows the audience to identify which object the stimulus was presented with.

It seems to me, moreover, that similar considerations will hold in other cases of moderately abstract explanations that on the surface appear to be penalized by **Specificity**. For instance, it is clear that if we consider the explanation of the window-breaking (4), we can tell a story similar to the one just told to explain why it doesn't count against the explanation of the window breaking (4) that it doesn't tell us precisely which rock was thrown at the window. Overall, then, the worry that **Specificity** might `undershoot' in an objectionable way is not one that should keep interventionists up at night.

A reviewer suggested another example (a variation on the Königsberg bridge problem) where **Specificity** may appear to give the wrong results. Four islands are connected by a system of bridges. Variables D_1-D_4 represent the degrees of the islands: that is, whether there is an odd or even number of bridges departing from them. (D_i=1 if island i is odd-degreed and 0 otherwise). Suppose that D_1 and D_2 both take value 1 while the others take value 0. Suppose we want to

[29] Alternatively, one might hold that the phrase `a red stimulus' is an existential generalization, but that it is *used* by the speaker to refer to the particular stimulus with which it was presented, so that the phenomenon in question is pragmatic rather than semantic (see Lewis [1979], p. 350). We need not settle this issue here.

[30] Likewise, *Red*=0 should be read as equivalent to *Tile*=0 – that is, as representing the tile being non-red.

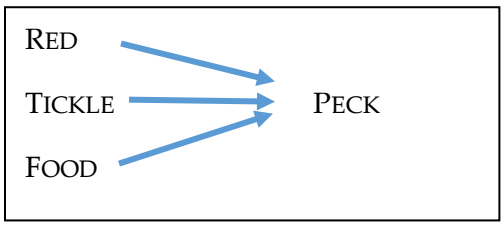
explain why it is impossible to trace an Eulerian circuit through the bridge system; that is, to cross each bridge exactly once and returning to one's starting point. The reviewer noted that it seems perfectly acceptable to explain this fact in terms of the theorem that an Eulerian path through a graph with at least one odd-degreed node is impossible and the fact that at least one of the islands is odd-degreed – in other words, in terms of the fact that D=1 (where D=1 if at least one of the four islands is odd-degreed and 0 otherwise). Yet D=1 is equivalent to $D_1=1 \vee D_2=1 \vee D_3=1 \vee D_4=1$. But in the present case altering the degree of D_3 or D_4 would clearly not make the system tourable via an Eulerian circuit. **Specificity** thus tells us that this explanation is not specific enough, and that one should also specify which of the four nodes were actually odd-degreed.

Two remarks are in order here. First, it is important to distinguish two different facts one might want to explain. First, one might be interested in understanding why certain systems can be toured via an Eulerian circuit while others are not. This, I take it, is the kind of general question that mathematicians are primarily interested. In this case, it is perfectly acceptable to answer the questions by pointing out that a system is untourable via an Eulerian circuit just in case it has an odd-degreed node. (This kind of `general explanation' doesn't fall under the purview of this paper, so **Specificity** doesn't rule against it.) On the other hand, one might be interested in understanding why *this* specific system cannot be toured via an Eulerian path. And in this case, it does seem to me that merely pointing out that the system has at least one odd-degreed node is not fully specific enough, and that one should also want to know *which* specific nodes are odd-degreed. As a case in point, imagine someone who knows that the specific bridge system of Königsberg has two odd-degreed nodes but is unable to identify which of the four landmasses are the nodes in question. (Perhaps for some strange reason that person is unable to count the number of bridges departing from each landmass.) I, at least, would be inclined to say that this person doesn't fully understand why Konigsberg is not tourable via an Eulerian circuit.

Second: consider an explanation of why no Eulerian circuit is possible in our system of islands that merely points out that D_1 and D_2 are odd-degreed. If this is all the explanation contains, it is in certain respects misleading: it suggests that any intervention making those two nodes even-degreed would make the system tourable, whereas this is clearly not the case. (An intervention that makes those nodes even-degreed by (say) removing one bridge from each of the first two islands may well leave the system untourable via an Eulerian circuit as removing those bridges may well make D_3 and D_4 odd-degreed.) To avoid such misleading implications, the explanation will have to include the claim that the system would still remain untourable if any of its nodes were odd-degreed (that is, if D took value 1). This makes for an important difference between this case and the case of (2) vs. (3). In the latter example, **Specificity** tells us that we can and should simply dispense with (3) and replace it with another explanation that invokes a different variable. In the present case, however, while **Specificity** tells us that merely mentioning that the outcome is due to the fact D took value 1 isn't a fully specific enough explanation, the consequence isn't that we should simply dispense with that explanation, but that we should simply complete it by specifying which nodes were odd-numbered. The fact that the outcome occurs as long as D takes value 1 remains an indispensable part of the explanatory story, although not the whole story.

To close this section, let me mention some recent considerations put forward by Woodward (unpublished, sect. 4) in response to Franklin-Hall and briefly compare them with the

account offered in this section. Woodward argues, like I did, that a variable such as *Full* is in certain crucial respects inadequate, although for reasons different than the ones I offered. His focus is on type-level causal relationships – the sorts of causal relationships in which scientists are ordinarily interested. In our case, the factors that cause peckings may be represented either via three different variables RED, TICKLE and FOOD, or via a single variable FULL lumping these factors together.[31] But as Woodward points out, from a methodological point of view the first choice of variables is superior to the second. For instance, the first choice of variables allows us to draw the following causal graph:



And given the way such graphs are conventionally interpreted, this one tells us (correctly) that if for instance the pathway from RED to PECK were `cancelled' or `perturbed' (for example by permanently blindfolding the pigeon), this would not affect the other paths present in the graph, so that it would still be possible to control PECK by intervening on TICKLE or FOOD. On the other hand, choosing a single variable FULL yields a graph containing a single arrow from FULL to PECK that doesn't contain this information. Likewise, the choice to represent the relevant causal factors via separate variables carries with it the (in this case correct) implication that those factors can be manipulated independently of one another, a fact that the second choice doesn't make clear. In these and other ways, Woodward argues, the first choice of variables allows us to reveal important facts about the causal structure of the `system' under consideration that the second fails to capture, and is therefore methodologically sounder.

A full comparison of my and Woodward's response to Franklin-Hall must await another occasion, but let me say that in my view these considerations are entirely right, and fully compatible with the account of the defectiveness of *Full* offered in this section. In different ways, both accounts show that the fact that *Full* disjunctively `lumps together' three variables that are distinct from one another makes that variable inadequate: Woodward puts the accent primarily on the fact that such a lumping is detrimental to the investigation of type-level causal relationships, my account focuses on the fact that it is detrimental to what interventionists regard as the goal of singular explanation. In that respect, my account nicely complements Woodward's considerations. Although Woodward doesn't discuss the case of singular explanation in detail, his suggestion seems to be that a singular explanation involving *Full* is inadequate because the type-level equivalent of this variable is methodologically defective. But if my argument is correct

[31] I use small caps for these variables to highlight the fact that by contrast to the variables used previously in the paper they are intended to represent general factors rather than singular events or circumstances. RED=1, for instance, represents the event-type `presentation of a red stimulus' rather than the colour of the specific stimulus with which the pigeon was presented in the particular circumstance that (1), (2) and (3) aim to explain.

there is another powerful and illuminating way for the interventionist to explain why an explanation such as (3) is defective: it does a less than satisfactory job at satisfying what interventionists regard as the fundamental goal of singular explanation. In this way, my solution highlights how elegantly interventionism handles the Goldilocks problem, by offering a unified explanation of the superiority of moderately abstract explanations over inappropriately detailed and overly abstract ones: in both cases, the superiority of moderately abstract explanations can be tied to the fact that they do a better job at identifying interventions under which the explanandum would have been different.

4 Conclusion

Much recent work on explanation – most notably on explanation in the special sciences– emphasizes the explanatory value of abstracting away from idiosyncratic details of the case at hand in order to highlight broad causal-explanatory patterns. Yet, as the comparison between explanations such as (2) and (3) reveals, there is undeniably something right to Salmon's ([1984]) and Lewis's ([1986a]) claim that good explanations should also be detailed enough in order to perspicuously locate their explananda within the causal structure of the world. If the arguments presented above are correct, the interventionist account of explanation is able to do justice to both of those lines of thought and to show that they are fully compatible with one another. More precisely, as section III showed, interventionism captures Salmon's and Lewis's intuition about the value of specificity, in the form of a privilege for explanations that zoom in on exactly those factors that could be intervened upon so as to change the explanandum, and ignore other distinct factors that could not be intervened upon in such a way. But good explanations should also exhaustively and perspicuously describe the pattern of dependence of the outcome on its causes – that is, the full and exact range of interventions on those factors that are associated with a change in the outcome. And as our examination of proportionality revealed, this goal is better satisfied by explanations that filter out certain details – specifically, explanations that do not distinguish between various alternative ways in which the cause might have brought its effect. From an interventionist picture, then, it is no wonder that we regard moderately abstract explanations as optimal, as such explanations display both the kind of specificity and the kind of abstraction that are valuable for the task of identifying explanandum-changing interventions.

Acknowledgements

I would like to thank Max Kistler, James Woodward and two anonymous referees for very valuable comments.

Department of Philosophy
Illinois Wesleyan University
Bloomington, IL 61701, USA
tblancha@iwu.edu

References

Fodor, J. [1974]: 'Special Sciences', *Synthese*, **28**, pp. 97–115.

Franklin-Hall, L. [2016]: `High-Level Explanation and the Interventionist's "Variables Problem"', *British Jour nal for the Philosophy of Science*, **67**, pp. 553–577.

Garfinkel, A. [1981]: *Forms of Explanation*, New Haven, CT: Yale University Press.

Halpern, J. and Hitchcock, C. [2010]: `Actual Causation and the Art of Modeling', in R. Dechter, H. Geffner & J. Halpern (*eds*), *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, London: College Publications, pp. 383–406.

Halpern, J. & Hitchcock, C. [2015]: `Graded Causation and Defaults', *British Journal for the Philosophy of Science*, **66**, pp. 413–457.

Hitchcock, C. [2001]: `The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy*, **98**, pp. 273–99.

Hitchcock, C. [2007]: `Preemption, Prevention and the Principle of Sufficient Reason', *Philosophical Review*, **116**, pp. 495–532.

Lewis, D. [1979]: `Scorekeeping in a Language Game', *Journal of Philosophical Logic*, **8**, pp. 339–59.

Lewis, D. [1983]: `New Work for a Theory of Universals', *Australasian Journal of Philosophy*, **61**, pp. 343–77.

Lewis, D. [1986a]: `Causal Explanation', in *Philosophical Papers*, Volume II, Oxford: Oxford University Press, pp. 214–240.

Lewis, D. [1986b]: `Events', in *Philosophical Papers,* Volume II, Oxford: Oxford University Press, pp. 241–269.

Loewer, B. [1996]: `Humean Supervenience', *Philosophical Topics*, **24**, pp. 101–127.

Salmon, W. [1984]: *Scientific Explanation and the Causal Structure of the World*, Princeton: Princeton University Press.

Shapiro, L. and Sober, E. [2012]: `Against Proportionality', *Analysis*, **72**, pp. 89–93.

Strevens, M. [2008]: *Depth*, Cambridge, MA: Harvard University Press.

Weatherson, B. [2012]: `Explanation, Idealisation, and the Goldilocks Problem', *Philosophy and Phenomenological Research*, **84**, pp. 461–73.

Weslake, B. [2010]: `Explanatory Depth', *Philosophy of Science*, **77**, pp. 273–94.

Weslake, B. [forthcoming]: `A Partial Theory of Actual Causation', *British Journal for the Philosophy of Science*, forthcoming.
Woodward, J. [2003]: *Making Things Happen*, Oxford: Oxford University Press.

Woodward, J. [2008]: `Mental Causation and Neural Mechanisms', in J. Hohwy and J. Kallestrup (*eds*), *Being Reduced*, Oxford: Oxford University Press, pp. 218–62.

Woodward, J. [2010]: `Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation', *Biology and Philosophy*, **25**, pp. 287–318.

Woodward, J. [2015]: `Interventionism and Causal Exclusion', *Philosophy and Phenomenological Research*, **91**, pp. 303 –347.

Woodward, J. [2016]: `The Problem of Variable Choice', *Synthese*, **193**, pp. 1047–1072.

Woodward, J. [unpublished]: `Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance'.

Yablo, S. [1992]: `Mental Causation', *Philosophical Review*, **101**, pp. 245 –80.