Summer 2015

Backwards Explanation and Unification

Richard Fry
Southern Illinois University Edwardsville, rfry@siue.edu

Penultimate draft.
Please cite the published version:
Fry, R. (2015). 'Backwards Explanation and Unification.' *European Journal for Philosophy of Science*. DOI: 10.1007/s13194-015-0121-1

Richard Fry
Southern Illinois University Edwardsville
rfry@siue.edu

Backwards Explanation and Unification

Abstract: It is an open question whether we ever successfully explain earlier states by appealing to later ones, and, further, whether this is even possible. Typically, these two questions are answered in the same way: if we give and accept 'backward explanations,' they must be possible; if they are impossible, we are right to reject them. I argue that backwards explanations are brittle—they fail if the future event does not occur—and this is part of the reason they are not accepted about the actual world. This does not mean, however, that they must be rejected entirely. I argue that backwards explanations are possible for certain systems. This shields unificationism about scientific explanation from some recent criticisms.

1. The Questions of Backwards Explanation

Can anything be explained by appealing to something that is temporally later? There are two related issues. First, are these 'backwards explanations' given and accepted? Second, could they ever be successful, and under what conditions? Giving answers to these questions shows us some constraints on theories of explanation.

Jenkins and Nolan (2008), for instance, answer both questions affirmatively, arguing that backwards explanations are common. By contrast, Woodward (2003) should be seen as arguing that the answer to both questions is 'no.' Woodward claims that backwards explanations are neither given nor accepted about the world we live in. There are two ways of understanding this argument: it might be trading on our explanatory intuitions about the

actual world or our explanatory intuitions more generally. In either case, I argue, Woodward is pushed to make the strong claim that backwards explanations are in principle unsuccessful, that is, that they are never acceptable for any sort of system.

There are real consequences to this rejection of backwards explanation: Woodward has argued from the fact that unificationism does not rule out backwards explanations for certain systems to the conclusion that it is a non-starter as a theory of scientific explanation. Instead, he claims, a theory of explanation must countenance only explanations that run the same direction as causation within the system that they are about. Woodward's concerns about unificationism are caught up in his views about causation, but the argument against unificationism rests on claims about explanation more generally. But it seems too quick to reject unificationism about scientific explanation and draw conclusions about causation on the basis of a tendentious claim about backwards explanations.

I will argue that the two questions of backwards explanation should be answered 'no' and 'yes,' respectively: we do indeed reject backwards explanations of phenomena in the world in which we live, but backwards explanations can still be successful for certain sorts of systems, namely systems genuinely indifferent to the direction of time. This counts in favor of unificationism, as unificationism can accommodate the fact that backwards explanations are not given or accepted about the actual world but still potentially successful for some systems.

2. Backwards Explanations Not Common

Using temporally subsequent states to explain temporally prior ones is to offer a 'backwards explanation.' Jenkins and Nolan (2008) argue that we regularly give and accept backwards explanations. One key case they consider is the scarlet pimpernel: the scarlet pimpernel is a flower that closes its petals before it rains. Thus, Jenkins and Nolan say, it is reasonable to explain the closing of the pedals in terms of the future rain.

(1) "Its flowers are closing because it is going to rain." (2008: 109)

This is, they say, a successful case of explanation wherein the explanandum is temporally prior to the explanans. They find other, similar patterns of explanation (2008: 104):

(2) "I'm tidying my flat today because my brother is coming to visit tomorrow."

(3) "The planet is slowing down because it is going to reach its apogee soon."

(4) "The volcano is smoking because it is going to erupt soon."

Jenkins and Nolan argue that it is not possible to reinterpret all of these cases as explanations of later states in terms of prior ones. They claim it is inappropriate to recast the pimpernel explanation, which was originally given in terms of future rain, in terms of the mechanisms that cause the pimpernel to close in response to changes in lighting. They take it that the person offering the explanation will often be ignorant that it is indeed changes in lighting that prompt the pimpernel to close. They say that it is problematic to systematically attribute to an explainer claims about mechanisms unknown to the explainer (2008: 108).

Byerly (2012) disagrees, citing the standard philosophical practice of reinterpreting a speaker's meaning, even in accordance with theories that they do not know of or endorse. Byerly cites Merricks (2001) as a key example; Merricks maintains both a minimal ontology and the truth of the statements about objects by re-interpreting claims about X's (tables, chairs, planets, *etc.*) in terms of 'simples arranged X-wise.' Byerly claims that we can reinterpret backwards explanations in whatever way we see fit because we have a legitimate philosophical practice that allows such reinterpreting.

But Byerly's exoneration of extreme reinterpretation is too quick. Merricks's interpretational scheme is supposed to be neutral (in some sense) with respect to the content; it is simple mechanical substitution, replacing every instance of 'table' with 'simples arranged table-wise,' *etc.* Merricks's scheme is only plausible if it does not really change what is said. If the speaker did not grant that the table just is some simples arranged table-wise, it would be inappropriate to reinterpret their speech in this way. Indeed, on Merricks's own view, talk of 'persons' cannot be reinterpreted as talk about 'simples arranged person-wise,' because persons are something over and above the collection of simples that make them up.

So while we can reinterpret explanations, we cannot change what is said *too much*. In the case of explanation, substituting mechanism talk for future event talk is a substantive change: one can believe that the claim about the future is explanatory while rejecting any description of the mechanism actuates the event—or even rejecting mechanisms entirely! Thus the substitution seems to go too far. But there are some cases where we would want to reinterpret a speakers claim about future events explaining present ones. Cases where charity

requires it, for instance. Consider the following case, where Annalise and Beatriz discuss the pimpernel:

A: Its flowers are closing because it's going to rain.

B: But it's not going to rain.

If Beatriz is right and we take Annalise's claim at face value, Annalise has failed to explain the pimpernel's is closing. Annalise's explanation appeals to a future state of affairs that will not come to pass, and while we might want to allow for some cases of false statements explaining, Annalise's appeal does not seem to explain. This suggests that backwards explanations only succeed when the future state they appeal to will actually occur.

We can protect seemingly backwards explanations from failing this condition by recasting them in terms of present circumstances rather than future ones, though. Take, for instance, this plausible extension of the scenario:

A: Its flowers are closing because it's going to rain.

B: But it's not going to rain.

A: Yes, but the pimpernel *thinks* it's going to rain.

Annalise recasts her initial explanation, saying that what she should have said is that its flowers are closing because the pimpernel expects rain. This is a way of reinterpreting Annalise's initial claim in a way that does minimal violence to it while still shifting it from a backwards explanation to a forwards one. We should see Annalise as offering (1A) from the beginning:

(1A) Its flowers are closing because it thinks it's going to rain.

Jenkins and Nolan claim that there is no reason to see this forwards explanation as the 'real' explanation (106). But the brittleness of backwards explanations does give us a reason: we can make these explanations more robust in light of uncertainty about the future by recasting them in terms of present expectations. Nothing in the present expectation of a future event depends on that future event's being actual, so the explanation can succeed even if that event does not come to pass. By shifting the claims about the future behind the propositional-attitude indicator, we have successfully dispensed with the need for the future state to be actual. The flat-tidying case can be dealt with similarly:

(2A) I'm tidying my flat today because I believe my brother is coming to visit tomorrow.

The cases are still different, though: we might be unhappy letting the intentional explanation stand in the pimpernel case. Because we think pimpernels are not intentional systems, we might want to reduce the 'thinks' talk to mechanism talk. That we are willing to accept one intentional explanation but not the other reflects our desire for explanations to match the systems they are given about, that is, only attribute systematic features to those systems that they actually have.

In this case, the intentional explanation 'matches' the system that it is about in the flat tidying case, because I am an intentional system. The intentional explanation does not match the system it is about in the pimpernel case, because the pimpernel is not. We can allow that I think, but claiming that the pimpernel thinks seems problematically false. Whatever else these explanations claim about the systems involved, true or false, they only seem acceptable when they appropriately take account of certain key characteristics of that system.

We might, then, end up recasting the intention talk in the pimpernel case as mechanism talk. Notice that it is only this last reframing that falls afoul of Jenkins and Nolan's objection. The issue they identify does not arise when we replace the appeals to future states with appeals to present ones; it only arises when we recast intention talk in terms of mechanisms, as it is only this last reframing that is not plausibly available to Annalise. Thus the problem is not with replacing talk of future states with present ones but with replacing intention talk with mechanism talk. As such it is a problem for the theorist of mind, not the theorist of explanation.

Regardless of whether we remove the intention talk or not in these cases, making them explicitly intentional made them significantly more robust and did so without falling afoul of Jenkins and Nolan's worries.

It would seem that some cases of backwards explanation cannot be dispensed with in this way, though. Cases that do not involve agents (or pseudo-agents) responding to the environment cannot be recast in these terms without changing what is meant too much. Cases like (3) "the planet is slowing down because it will reach its apogee soon" and (4) "the volcano is smoking because it will erupt soon" show that recasting all backwards explanations as intentional will not work.

As Jenkins and Nolan note, both of these cases depend on a regularity between what is happening now and what will happen in the future; the speaker appeals to some phenomenon regularly connected but temporally subsequent to the explanandum as explanatory of it. It appears the bare regularity is responsible for the explanation's success, and this essentially requires an appeal to a future state.

The volcano case is still brittle, though: if the volcano does not in fact erupt, its (future) erupting cannot explain its (present) smoking. And this shows that even regularity explanations require the future to cooperate. But they do seem to leave open the possibility of genuine backwards explanation, when the future does indeed conform to our expectation. This leads to a puzzle, though. An explanation, seemingly, should succeed or fail when it is given, not in light of what eventually occurs. This feature of explanation is part of what made the pimpernel case so clear cut. Beatriz denied *right then* that the future event would occur, updating the context so as to make conversational moves that depended on that claim unwarranted. Beatriz's claim narrows the range of possible explanations by explicitly excluding all appeals to that particular future event. It also seems to implicitly narrow the range of possible explanation schema to those which do not rely on *any* future event, because it makes the epistemic risk associated with any backwards explanation real. This implicit updating reinforces my claim that genuinely backwards explanations are avoided, when possible, because they are brittle.

This line of reasoning also explains why the denial that the planet will reach its apogee does not seem to undercut the explanation offered: the regularity between the planet's slowing and its approaching its apogee is so tight as to be nearly necessary. Even if something were to keep the planet from reaching apogee, the explanation still seems successful. Note that this non-brittle backwards explanation is strange in light of the others. That the apogee explanation does not fail when the claim about the future is false is not good reason to think that there are cases of backwards explanation that do not fail when their claims about the future are false. Rather, it is reason to see the apogee case as a-typical

of backwards explanation. This is because it is not a case of backwards explanation at all. To see why, consider the following two versions of the explanation:

(3) "The planet is slowing down because it is going to reach its apogee soon."

(3A) "The planet is slowing down because it is approaching its apogee."

To most readers, these two explanations say the same thing; their content is the same in most conversational contexts. Further, if a speaker believes (3), they almost certainly believe (3A). This means we can unproblematically substitute (3A) for (3) in almost all conversational contexts. Our unwillingness to reject (3) when confronted with its falsity, then, almost certainly stems from the fact that we read it as actually asserting (3A). Note that (3A) is a forwards explanation, and it is still true even if the planet never reaches apogee.

This leaves only the volcano case, which relies fundamentally on a regularity (so resists reformulation) but is still brittle. Jenkins and Nolan consider recasting regularity explanations like this as calling on present dispositions instead of future events, changing (4) into (4A):

(4) "The volcano is smoking because it is going to erupt soon."

(4A) "The volcano is smoking because it is disposed to erupt."

They reject this substitution because the reformulation offered is strictly speaking stronger than the explanation it replaces. But this is exactly what we are looking for: if we understand the original explanation as genuinely making a claim about the future, then it is brittle. In order to make these explanations more robust in the face of uncertainty about that future, we do indeed want a stronger claim than the original explanation made. Their considerations do not give a reason to reject the reformulation in terms of dispositions but

rather to accept it. Alternatively, we might think that people commonly use locutions like (4) to mean (4A). If that is the case here, then there is no problem in making the substitution.

One might worry that this is like recasting the pimpernel explanation with mechanisms: the person offering (4) might have no knowledge of the particular properties that underwrite the disposition named in (4A). This worry is erroneous. The explainer need not have any particular knowledge about these dispositions in order to name them by their most common, most important effect. The volcano explanation can be unproblematically recast as appealing to dispositions.

On the other hand, one might worry that someone who understands a system well can know that some future event will happen, making a backwards explanation based on that knowledge successful. Consider a trained volcanologist: when the volcanologist has appropriate information, it would seem that they know the volcano will erupt and can use this as a way of explaining the present behavior of the volcano. But note that even Jenkins and Nolan allow recasting experts as offering forwards explanations; they objected to reinterpreting non-expert speakers who are ignorant of the underlying mechanisms as if they were experts. Here, the speaker's knowledge of the mechanisms involved underwrites both the claim about the future and the claim about the link between the present event and the future one. For experts, their knowledge of the underlying features of the system are the

root of both the predictive claim and the explanatory one, so the explanation can be easily understood as going from temporally earlier to later.[1]

The brittleness of backwards explanations pushes us to reframe them as forwards explanations. Responsiveness is explained intentionally, and bare regularities are explained with dispositions. Jenkins and Nolan's cases can all be recast as forwards explanation without doing violence to the claims made. This leads to the conclusion that backwards explanations are not common, though superficially backwards explanations surely are. We can make sense of more of our explanatory practices—namely that backwards explanations are brittle and that explanations succeed and fail when given—by understanding them as forwards explanations, and charity often requires this.

3. The Argument Against Any Backwards Explanation

Jenkins and Nolan fail to give any convincing cases of backwards explanation. This suggests that the right answer to 'Are backwards explanations actually given and accepted?' is 'No.' Seeing that backwards explanations are not accepted in our every day lives, Woodward makes the broader claim that backwards explanations violate some intuition we have about the nature of explanation. In short, particular backwards explanations are not successful because backwards explanations are, broadly speaking, unacceptable to us. Woodward's

[1] Further, practicing volcanologists seem only to make claims about the likelihood of an eruption (of a particular sort) over some time-frame (Fountain 2015). Their predictions seem to be sensitive to uncertainty about the future. It would then be appropriate to treat their explanations as being similarly sensitive, that is, as being given in terms of the information they have and not the predictions they are making.

discussion of backwards explanation is a key part of his argument against unificationism about scientific explanation.

Kitcher (1989) posits that scientific explanations explain by unifying.[2] Scientific explanations unify phenomena by showing how some of them—the *explananda*—are derivable from the others. This derivation is the *explanans*. These derivations are explanatory when they adhere to a schematic form found in the culturally-specific explanatory store. For the derivation to be explanatory, the explanatory store must consist of whatever set of schematic sentences can explain all the phenomena with a minimum of derivation patterns, i.e., it must be maximally unified. For instance, because I can explain the dimensions of *any* object by its origin and development but only *some* objects by the size of their shadows, derivations of sizes from shadows are non-explanatory, while origin and development derivations are.[3]

Woodward (2003) wants to show that Kitcher's view that causal judgements are parasitic upon and can be recreated from explanatory practices is false. Kitcher claims, in part, that causal judgements result from and track our explanatory practices. Woodward wishes to show the opposite, that causal judgements precede explanation and enable it. To show this, Woodward argues that Kitcher's unificationism cannot recreate the time asymmetry we see in our explanatory practices, thus the time asymmetry that is evident in

---

[2] Friedman (1974) first posited this view, but his version faced several objections—see Kitcher (1976) for a discussion. Kitcher developed his version to fix some of these problems.

[3] The D-N model, as found in e.g., Hempel and Oppenheim (1948), suffered from this derivation-of-dimensions-from-shadows problem. The problem was originated by Bromberger and is found in modified form elsewhere. See Salmon (1989: 47n12) for a discussion.

our causal judgements must be primary or fundamental, meaning that causation is the more fundamental notion.

To argue that the unificationist cannot recreate the time asymmetry of our causal judgements from our explanatory practices (as described by unificationism), Woodward, following Barnes (1992), asks us to consider a closed system governed by laws indifferent to the direction of time, a solar system governed by Newtonian mechanics.[4] Those laws, Woodward stipulates, are time-direction indifferent: they can be used equally well, given a complete accounting of one particular state of the system as an input, to derive both later and earlier states of the system. For this system, the *predictive* set of derivations will be no more unifying than the *retrodictive* set of derivations, so a maximally unified explanatory store could feature either on Kitcher's account, Woodward claims.

But, Woodward notes, we do not accept backwards explanations for this system. He says:

> However, we ordinarily think of the predictive derivations and not the retrodictive derivations as explanatory and the present state of the planets as the cause of their future state and not vice-versa. (2003: 362)

Here he is following Barnes, who says a bit more:

> The Newtonian Predictive Pattern is intuitively explanatory of the members [of the set of statements describing the velocities and positions of objects in the system]. A perfectly legitimate explanation of the fact that some object has a particular position and velocity at some time would consist of a citation of the fact that the system containing the object had a particular state at some earlier time, together with the deterministic Newtonian laws that led inexorably from the latter to the former. However, the Newtonian Retrodictive Pattern is utterly

[4] I follow the Woodward development of this objection, and have modified the exposition to make the case clearer, more consistent and simpler to explain. There are two ways of interpreting the objection, and I address both of them in turn.

> nonexplanatory of the members of [the set]; such explananda cannot in general be explained by citing facts that occurred subsequent to the explananda themselves. (1992: 565-6)

Woodward and Barnes claim here that, in this case at least, predictions explain and retrodictions do not. No explicit argument is provided for the claim that the retrodictive pattern, applied to this system, is non-explanatory. Indeed, both authors have stipulated that the laws are sufficient in this system to produce accurate derivations of any state from any other state. Because the laws are sufficient to accurately predict and retrodict *ex hypothesi*, it cannot be the case that the retrodictive pattern is non-explanatory because it is insufficient to produce accurate accountings of the other states of the system. What remains is its character as backwards explanation rather than forwards explanation: indeed, for the patterns to be equally unifying—and thus for the system to be apt for critiquing the unificationist—the only difference between the forwards and backwards explanations must be their direction. The reason the backwards explanations are non-explanatory, then, is because they are backwards explanations. Woodward and Barnes conclude that unificationism fails as a model for explanation, because this case shows that backwards explanations are unacceptable and unificationism does not exclude retrodictive patterns from the explanatory store.

This objection is in need of clarification along three dimensions. First, Woodward cannot be claiming that, in order to be explanatory, the explanandum needs to be in the future when the explanation is given and the explanans in the past. If this were Woodward's demand, then no prediction where both of the events are in the past (or future) from the perspective of evaluation would be explanatory, even if the prediction preceded the event

predicted. But it seems like there are genuine cases of explanation where both the act of prediction and the event predicted are now in the past from our perspective (e.g., predicting past eclipses before they occurred).

Second, Woodward's claim cannot be about 'prediction' and 'retrodiction' in the normal sense of those words.[5] Taking two events that are already in the past when the explanation is given and deriving the later one from the earlier one can be genuinely explanatory even though it is not a 'prediction' in the normal sense (e.g., explaining past eclipses by the positions of celestial bodies leading up to them).

This means that Woodward's claim must be understood in terms of a temporal *sequence*, then, and not in terms of past and future; a predictive derivation, then, must be a derivation of a later state based on some earlier state. Thus clarified, the claim is that when giving explanations, the event or state appealed to in the explanandum must be temporally prior to the event or state that constitutes part of the explanans. Explanations of the state of the system at t_0 must always be given in terms of the state of the system at t_{-n}, never t_{+n}. Woodward's implicit claim is that (S1) is supposed to be explanatory, while (S2) is not:

(S1) "Because bodies B have locations C and momenta M (i.e., the system is in state R) at t_0 and the system is governed by laws L, they *will have* these other locations C_1 and momenta M_1 (i.e., the system will be in state S) later at t_{+1}."

(S2) "Because bodies B have locations C and momenta M (i.e., the system is in state R) at t_0 and the system is governed by laws L, they *did have* these other locations C_{-1} and momenta M_{-1} (i.e., the system was in state Q) earlier at t_{-1}."

Woodward's argument depends on the claim that, at least for this system, one state's being temporally prior to another state gives it special explanatory status with respect to the

[5] *Modulo* that I am not sure that 'retrodiction' has a 'normal sense.'

later state; as such, prior states can serve as the basis for an explanation of later states but later states do not serve as the explanatory base for prior states.

One final point of clarification is needed. Woodward's exegesis—and, in following him, the exegesis I have given above—is ambiguous between two different arguments. The first way of interpreting Woodward is as saying, of our actual solar system, that it has certain features that should make it as amenable to derivations of later states from prior ones as to prior states from later ones. That we do not in fact accept backwards explanations of our actual solar system despite its amenability to them, then, should lead us to conclude that unificationism is faulty because it does not rule them out.[6] The second interpretation of the argument posits a hypothetical system with laws indifferent to the direction of time and asks what our explanatory intuitions would be about that system. It then claims that because we would think backwards explanations unacceptable, even for this system that is maximally amenable to them, unificationism is incorrect because it does not rule them out.

One way of escaping the criticisms I will level at the first version of the argument is to reformulate the objection in terms of the second version of the argument, so I will address them in this order, ultimately showing that they are both untenable.

4. Actual World Interpretation of Woodward's Argument

The first version of Woodward's argument addresses the actual world, making the claim that we do not accept backwards explanations about certain actual world systems that would seem amenable to them. The argument runs like this: Consider our solar system as a

---

[6] I would like to thank an anonymous referee for alerting me to this reading of Woodward's concerns.

closed, Newtonian system. Despite the fact that such systems are governed by laws indifferent to the direction of time—that is, laws that can be used equally well to produce derivations of future states from prior ones as to produce derivations of prior states from future ones—observe that we accept only derivations of later states from prior states as being explanatory. Because the laws are indifferent to the direction of time, though, the forward-derivational pattern and the backwards-derivational pattern are equally unifying; thus unificationism cannot distinguish between them for this system. Because unificationism cannot rule out the use of later states to explain prior ones for this actual world system and because our actual explanatory practices disallow such explanations, we should reject unificationism (as it cannot generate the requisite temporal asymmetry). Note that Woodward's argument here does not depend on backwards explanations being in principle unacceptable, only the comparatively weaker claim that they are in fact not accepted for our solar system.

In evaluating this objection, it should be noted that our actual solar system is neither closed nor Newtonian. That is, the energy/matter total within the system is not fixed (electromagnetic radiation, for instance, reaches us from outside the solar system). Nor is it the case that Newtonian laws are completely accurate in accounting for the positions and momenta of the planets (consider, e.g., the precession of the perihelion of Mercury).

So, what Woodward must mean by citing our actual solar system as an example of a closed, Newtonian system governed by laws indifferent to the direction of time is that it would seem that the positions and momenta of planets in our actual solar system are amenable to characterization by and explanation with Newtonian laws indifferent to the

direction of time (and that the energy transfer between the solar system and the universe at large is negligible for those purposes). Over historical time, our predictions and retrodictions remain empirically adequate despite the deviations of the actual system from the idealizations being considered.

For the actual solar system, characterized as such, we are able to perform both derivations of later states from prior ones and prior states from later ones: the laws allow this because they are indifferent to the direction of time. This is required for Woodward's argument: for the two derivation patterns to be equally unifying, the laws must be equally accurate for both prior-to-later and later-to-prior derivations. Both patterns will be equally empirically adequate for this system considered as such. (That is, if they vary from our observations, the predictive and retrodictive derivation patterns will vary equally much from those observations.)

Woodward has claimed that we in fact accept only prior-to-later derivations as explanatory, despite the system's amenability to both patterns. He concludes that unificationism is inadequate to capture the temporal asymmetry in our everyday explanatory practices. But unificationism can indeed make sense of the fact that we, as we are currently situated, make use of temporal information to rule out certain sorts of explanations of actual world systems. This introduces the temporal asymmetry that Woodward seeks.

To see why this is the case, first consider two ways in which a system might be considered to be time symmetric. First, a system that is time symmetric *with respect of laws* has laws that work equally well to produce derivations of later states from prior ones and *vice versa*. Second, a system that is time symmetric *with respect of phenomena* exhibits no differences

in behavior when the time direction is reversed (and other corresponding changes made). That is, no observation could count in favor of seeing time as appropriately oriented in one direction rather than the other. Consider an idealized, frictionless billiards table: there will be no grounds for selecting one time direction over the other for this system.

The actual world is time symmetric in the first sense but not the second. That is, it is time symmetric with respect of laws but not phenomena. The actual world exhibits a number of apparently time-irreversible behaviors. The most obvious of these are its entropic behaviors: the diffusion of gasses, melting, friction, *etc*. These are all processes that require a set direction of time for an adequate characterization. So while the system is governed by laws that are indifferent to the direction of time, any adequate prediction or retrodiction involving these phenomena requires the introduction of some auxiliary hypotheses (e.g., about boundary conditions) that include time-direction information.

Even if we know that the derivational patterns (including the auxiliary hypotheses) are equally empirically adequate, we also know that time direction information must figure into our predictions and retrodictions for phenomena in this world. Because they are equally empirically adequate, though, prior-to-later and later-to-prior derivation patterns (that appeal to time-direction information embedded in the auxiliary hypotheses) still appear to be equally unifying.

But they are not, really: when we explain the actual world, we take advantage of and draw on what we know about the actual world. Explaining phenomena in the actual world requires time direction information (because it is time asymmetric with respect of phenomena). Because we are meant to be explaining the actual world in this example, we

draw on our experiences of that world—experiences which present it to us as being time asymmetric—in determining what explanatory patterns are appropriate or inappropriate. Note that this is how the argument must work. Unless Woodward is claiming that backwards explanations are in principle unacceptable—a claim I will examine shortly—there must be something about the system (that is, about the actual world) that makes it so that we do not accept backwards explanations about it. The unificationist should identify the most reasonable candidate for this feature as being that the actual world presents itself to us as time asymmetric with respect of phenomena. Because it presents itself to us this way, this is the standard we use when evaluating it (despite the fact that the underlying laws themselves are time-direction indifferent). Unificationism *would* make the patterns of forwards and backwards explanation equally adequate if we considered the system as an idealization. But this is explicitly not the standard that we were asked to use; on this version of the argument we were explicitly asked to consider the system *as actual*. As instructed, we judged it to the standard we use when we judge explanations about the actual world; that standard takes the world as we find it.

Unificationism can give pride of place to forwards explanations in this case because the system under consideration is an actual world system: when considering actual world systems, our experiences of that world factor in to our decisions about which patterns of explanation to deploy. Our lived experience of the actual world displays a sharp prior-to-later orientation; these experiences can be easily unified with each other and with accounts of the system we are considering within an over-arching prior-to-later derivational scheme, but hardly at all within a scheme of later-to-prior derivations. Considering the system as an

actual world system provides a new and vivid set of phenomena for patterns of explanation to accommodate, and a large proportion of these new phenomena display the earlier-to-later asymmetry of our everyday experience. Prior-to-later patterns will be more unifying overall when the system in question is to be explicitly considered as part of an overarching system (the actual world) which includes this set of time-asymmetric experiences within it. That the explanatory practices we actually have and use are rooted in the consideration of phenomena as they appear to us (that is, as time asymmetric) and not the (time symmetric) laws should be no surprise.

Woodward claimed that, for the system described in the actual world, unificationism could not claim prior-to-later derivational patterns as more unifying. But unificationism can generate temporal asymmetry by noting that, when considering the actual world, the derivational patterns must unify our experiences of the (time asymmetric) phenomena within it as well. Predictive patterns will thus unify more completely than retrodictive ones.

One way to escape these criticisms is to claim that our rejection of backwards explanations for the actual world is really rooted in the unacceptability of backwards explanations more generally. This would deny the unificationist's appeal to our lived experience, but it would also require showing that our general explanatory intuitions never allow for backwards explanations. The second version of Woodward's argument attempts to do precisely this.

5. Cases Show Backwards Explanation Sometimes Acceptable

The other version of Woodward's argument considers our intuitions about a

hypothetical system, not an actual one. If one took Woodward to claim that we should consider the system as an idealization or a hypothetical, the fact that the actual world is entropic (and otherwise time asymmetric) would not matter when we consider our explanatory intuitions about the system in question.

Woodward pointed out that patterns of explanation running from temporally later to temporally earlier are just as unified as those that run in the opposite direction for a closed system governed by time-direction indifferent deterministic laws. Unificationism does not disallow these backwards explanations, so, Woodward argues, because we think backwards explanations are problematic, unificationism is problematic. Woodward's argument depends on the claim that backwards explanations are problematic when given about this system. Because, for this case, the ground of the claim that backwards explanations are unacceptable cannot be that they are unacceptable for real systems, the ground would have to be that they are unacceptable generally. By considering the hypothetical and related cases more closely, I show that we should only rule out backwards explanations for systems that are not amenable to them. That is, I will argue that Woodward's (illicit, too strong) rejection of backwards explanation arises from the (weaker, reasonable) constraint on explanation that explanations must match the system they are about.

To see that backwards explanations succeed for this system, when considered as hypothetical, suppose that we did not know the direction of time in the system. This can be the case even though we do know the sequence of the states of the system; i.e., we know that state R comes between states Q and S, but we do not know whether Q is the earliest state in this sequence or if S is the earliest. So we do not know if, moving from earliest to

latest, the states are Q-R-S or S-R-Q. Appeals to the states and the laws would seem to explain the system's transitions between these states, even in the absence of temporal priority information; that is, it seems that explanation is still possible in this system. Each of the other states of the system can serve as the basis for an explanation of R, as we are able to derive R from each of the other states. This is inconsistent with Woodward's claim of special explanatory status for states which are temporally prior: here we have no temporal priority, only temporal ordering.

In order to make the example work, Woodward has allowed that explanations in terms or derivations of states from other states via time indifferent laws in the presence of time *direction* information is explanatory. This case makes that last codicil seem unnecessary: for this system, where we do not have time direction information, explanation still seems possible. We can still tell a complete story (for this system) about how it develops from one stage to the next, and this is the essence of explanation in both cases. This suggests that Woodward's restriction of explanation to prior-to-later explanations is inconsistent with our intuitions about explanation.

A second way to see this inconsistency is to consider two systems with laws indifferent to the direction of time. Suppose that the two systems are identical except for the direction of time in that system. These systems will be mirrors of each other along the temporal dimension. If Woodward is right, pairs of identical states will have opposite directions of explanation. But this seems wrong: if the laws really are indifferent to the direction of time, then the fact that one state explains another in one of these systems should suffice to show that it will explain it in the other as well. In this case, being temporally

prior does not seem to grant the special explanatory status that Woodward claims. These examples should be enough to show that the temporal asymmetry condition on explanation that Woodward advances is ungrounded.

We might yet think that there is a fact of the matter as to which state of the system is earlier in the Q, R, S example, and that this will be discoverable by examining the other quantities in the system. This would allow us to determine which states are 'really' earlier, and so save the constraint. But if the system is governed by time symmetric laws then the other quantities can be defined equally well given either assignment of direction to time. Supposing that time has either direction cannot be superior to proposing the opposite assignment. 'Earlier' and 'later' have no 'natural' orientation in this system (unless we illicitly import some other information, e.g., by assimilating the boundary conditions of this system to the boundary conditions of our own world). The terms 'earlier' and 'later' will have no specification for this system unless we arbitrarily select an orientation for the time-like dimension. It is because derivations of states from other states still seem explanatory in this system that prior/later relationships seem not to be relevant to the explanatory potential of any of the stages of the system.

According to this version of Woodward, no state of the systems we are now considering could possibly be explanatory of any other state unless we were to decide which direction we should take time to flow. To see that this is absurd, consider any two states of the system, A and B. Given one (arbitrarily selected) time orientation, A will be explanatory of B because A will be earlier than B. On this orientation, B will not be—*and could not be*—explanatory of A. But on the other orientation, B will be explanatory of A and A will not be

—*and could not be*—explanatory of B. Woodward would bar us from explaining any state of this system in terms of any other because we lack temporal priority information (either because there is no such information or because we are in principle barred from accessing it). If we were to explain in this system, we would have to decide which state of the system to take as being earlier simply so that we can explain the 'later' states in terms of the 'earlier' ones. It would seem better, in this case, to accept that each of the states of the system is equally explanatory of each other state of the system—and genuinely or successfully so—regardless of whether they are in or could be put in an earlier/later relationship. Woodward's temporal asymmetry condition on explanation should be rejected.

6. Hypothetical Case Shows An Explanation Must Match Its System

Yet, as the discussion above shows, we are loathe to accept explanations of prior states in terms of later ones when they are given about the world we actually live in. This should not be surprising, though: we take it that the direction of time matters for the world we actually live in (unlike the hypothetical system we were asked to evaluate in the second version of the argument). The explanations we offer and accept in common practice are time asymmetric for just this reason: because we take the world to be time asymmetric, our explanatory schema are time asymmetric as well.

The proper constraint on explanation is not one that disallows backwards explanations *tout court*. Instead, it is one that demands that explanations be time asymmetric *because, in common practice, we take it that the world they are about is time asymmetric*. Woodward has taken a defeasible, common-sense constraint on the explanandum (that the system to be

explained is time asymmetric) and confused it for an indefeasible constraint on the explanans (that it be *incompatible* with time symmetry). This is possible—and indeed understandable—once we see the related, real constraint on the explanans: that the nature of the explanation must match the nature of the system being explained.

This constraint should look familiar; it is the same constraint that played a motivating role in Section 2 above, and it is implicitly at work in Section 4. So let's return to non-scientific explanation for a moment. When I give an explanation of why I am tidying my flat in terms of my brother's future visit, my successful explanation is given in terms of psychological states. When I explain the pimpernel's closing by appealing to psychological states, my explanation seems problematic. That is because we take it that I actually do have psychological states, unlike the pimpernel. I am a system that is apt to psychological explanation in a way that the pimpernel is not. The explanations need to match the system about which they are given in order to be successful. In the first case, this is accomplished by appealing to psychological states; in the second case it is done by not appealing to psychological states. The nature of the successful explanation tracks the sort of system being explained. When we take certain sorts of systems (the actual solar system, the pimpernel, me) to have certain sorts of features (being time asymmetric with respect of phenomena, being non-intentional, being intentional), the explanations given about those systems need to respect that those systems have those features. An explanation should be consistent with those sorts of systems having those sorts of features. This is what it means for an explanation to 'match' the system that it is about.

This new constraint—that explanations match the system about which they are given—shows why backwards explanations are both uncommon but genuinely successful in some cases. Backwards explanations are genuinely successful in some cases because some systems really are time-direction indifferent, that is, they are time symmetric in both senses given above. But we rarely give and accept backwards explanations because we rarely are explaining systems that we believe to be time-direction indifferent. The two questions of backwards explanation—do we give them, and are they ever successful—are thus best answered 'not really' and 'yes,' respectively.[7]

This analysis counts in favor of unificationism. Unificationism allows that if the world turns out to be time asymmetric, explanations of the world ought to be time asymmetric, too. *Whatever way the world is*, the explanations aimed at it ought to match the way that it is; this is not a revision to our common sense view of explanation, it *is* our common sense view. Unificationism allows us to accept backwards explanations for systems that really are time symmetric (in both senses), but to reject them in day-to-day cases of pimpernels and volcanos. There is nothing about unificationism that suggests that the explanatory store of a time asymmetric world ought to be populated with time symmetric explanatory frames that would fail in many application conditions. If the world turns out to be amenable to characterization with time symmetric resources, then the explanations aimed at that world

[7] One might worry that the 'backwards' in 'backwards explanation' makes no sense for a system that is time-direction indifferent, thinking that all explanations will be in some sense a-temporal and therefore not 'backwards.' This is not so, though: there can be time-direction indifferent systems that do indeed have a set direction of time, it will just be the case that the direction of time will not figure in explanations for that system. It is true that their 'backwardness' does not make these explanations as interesting in a time-direction indifferent system as they would be in a temporally asymmetric system, but they would still comprise a distinguishable class of successful explanation.

ought to evince that same time symmetry. Unificationism allows for the common-sense criterion that the explanation must match the system to be explained and can accommodate either of these scientific results. These are major virtues of unificationism.

7. Causation and Conclusion

One final point. As noted above, Woodward's attack on Kitcher and unificationism arises from a discussion of causation. Woodward's concern is to refute Kitcher's claim that causal judgements can be cashed out in terms of explanatory judgements. Woodward says:

> [Kitcher's] claim is that our ordinary judgments about causal asymmetries can be derived from the unificationist account. The example just described casts doubt on this claim. More generally, it casts doubt on Kitcher's contention that one can begin with the notion of explanatory unification, understood in a way that does not presuppose causal notions, and use it to derive the content of causal judgments. (2003: 362)

Woodward is right that the unificationist can only make causal judgements time asymmetric by making explanations time asymmetric as well, precisely because the unificationist seeks to produce causal judgements from explanatory ones. As we saw above, the unificationist can produce explanatory time asymmetry for the actual world by noting that our explanations of the actual world must be unified with our experiences of the phenomena in it and that those experiences take the phenomena to be time asymmetric.

Unificationism is thus well-situated to explain our common sense judgements about causation. Because the explanations we actually give and actually accept presuppose time asymmetry, so too will our causal judgements. Our explanatory store only countenances explanatory patterns for this world that run from earlier to later, so, causal judgements based on those explanations will evince a similar time asymmetry. The unificationist can parse

causal judgements entirely in terms of explanation and yet still retain the temporal asymmetry that Woodward sees as required for explanations of and causal judgements about this world. Unificationism is strongly consistent with the intuition that explanations must match the system that they are about, and once we have that constraint, unificationism is well-positioned to show how causal judgements are generated from explanations not just for this world but more generally.

Understanding backwards explanation has consequences for our understanding of scientific explanation generally and, more specifically, for the status of unificationism. It also has consequences for how we think about causation. Here I have shown that while backwards explanations are not (usually) in fact given and accepted about the world we live in, we should still take it that they could be successful (for certain sorts of system). For the reasons outlined above, this shows that unificationism is still a viable candidate for a theory of scientific explanation and can, perhaps, even explain how we come to make the causal judgements that we do.

References:

Friedman, M. (1974) 'Explanation and Scientific Understanding,' *The Journal of Philosophy* 71 (1), 5-19.

Barnes, E. (1992) 'Explanatory Unification and the Problem of Asymmetry,' *Philosophy of Science* 59 (4), 558-571.

Byerly, T.R. (2012) 'Explanationism and Beliefs About the Future,' *Erkenntnis* 78, 229-243.

Fountain, H. (2015) 'Pressure, and Mystery, on the Rise,' *The New York Times*, retrieved from http://www.nytimes.com/2015/01/06/science/predicting-what-a-volcano-may-or-may-not-do-is-as-tricky-as-it-is-crucial-as-iceland-well-knows.html

Jenkins, C.S. and D. Nolan. (2008) 'Backwards Explanation,' *Philosophical Studies* 140 (1), 103-115.

Kitcher, P. (1976) 'Explanation, Conjunction, and Unification,' *The Journal of Philosophy* 73 (8), 207-212

Kitcher, P. (1989) 'Explanatory Unification and the Causal Structure of the World,' in *Scientific Explanation*, P. Kitcher and W. Salmon (eds.), University of Minnesota Press, Minneapolis.

Merricks, T. (2001) *Objects and Persons*, Oxford University Press, Oxford.

Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press, Oxford.

Hempel, C. and P. Oppenheim. (1948) 'Studies in the Logic of Explanation,' *Philosophy of Science*, 15, 135—175.

Salmon, W. (1989) *Four Decades of Scientific Explanation*. University of Minnesota Press, Minneapolis.