Flagpoles Anyone? Causal and Explanatory Asymmetries[1]

Forthcoming: *Nowhere*

Note to reader: Because of its length (among other considerations) this paper is likely not publishable in any journal. Since it is now in somewhat stable form and I am not going to do anything more with it in the near future, I thought that I would post it on the Philsci-archive. At least in this way I can avoid all of the referees who will hate it.

James Woodward

HPS, Pittsburgh

1. Introduction

A long-standing puzzle in philosophy of science concerns the direction of explanation (and causation). As a familiar illustration, discussed by Hempel, 1965 suppose that we are given information about the height $H=h$ of a flagpole, the length $S=s$ of the shadow it casts on the ground (assumed to be level and at right angles to the pole) as a result of the light provided by the sun and the angle $A= a$ between the shadow and the sun. Then from the values of any two of these variables and laws concerning the rectilinear propagation of light we can derive or deduce the value of the third. None the less only one of these derivations (from H and A to S) is thought to be explanatory (or to track the direction of explanation)—a derivation of H from S and A is no explanation. What is the source of this asymmetry or directionality? Why do we regard one of these derivations as explanatory and the other as not? How can we tell whether we have got the direction of explanation right? Or is there even such a thing as an objectively correct direction in such cases?

---

[1] An early version of this paper was given as a talk at the "Hempel and Beyond" workshop at the University of Cologne in 2015 (that is part of the reason for the flagpoles in the title). I also gave versions as talks at the LSE conference in honor of John Worrall in 2016, at UC-Irvine, at CMU and to the HPASS reading group at UCLA. I am grateful to the audiences in all of those places for helpful comments.

I also especially want to thank a number of others who either commented on earlier versions of this paper or discussed its content (or ideas related to its content) with me. These include Matt Farr, Clark Glymour, Marc Lange, John Norton, Reuben Stern, Porter Williams, Kun Zhang, and Jiji Zhang.

The reader may well wonder why, with all of this illustrious help, this paper is not a lot better. This is a causal inference problem and the answer is the obvious one.

A very similar issue (arguably the same issue, at least insofar as our focus is on *causal* explanation) arises in connection with causal inference. Suppose that *X* and *Y* are correlated[2] random variables. Suppose that we can exclude the possibility of confounding or common causes, so the only two alternatives are that *X* causes *Y* or that *Y* causes *X*. Is there some way we can reliably infer, given other assumptions and perhaps information about other correlations (e.g., correlations involving a third variable *Z* with *X* and *Y*) whether the causal direction is from *X* to *Y* or from *Y* to *X*? Why, for that matter, do we think of causation as having directional or asymmetric features at all? What can we say about the source of these features? How do these relate to other features that we think that causal relations possess?

This essay explores some of these issues. For most of this paper by "explanation" I will mean causal explanation. The penultimate section (12) will consider the extent to which the framework I provide might be extended to asymmetries present in non-causal explanations. The background theory of causation I will assume is the interventionist theory described in Woodward (2003). For our purposes, we will need only the following simple version:

> (**M**) *X* causes *Y* if and only if (i) it is possible to intervene to change the value of *X* and (ii) under some such intervention on *X*, the value of *Y* would change

An intervention on *X* is an unconfounded manipulation of *X* that changes a second variable *Y*, if at all, only through the change in *X*. For present purposes we can think of it as broadly the same notion as is captured by Pearl's "do" operator[3]. (Pearl, 2000, 2009). An intervention on *X* can be "hard" or "arrow breaking" in which case it puts the variable intervened on entirely under the control of the intervention, breaking the connection with all other causes of *X*. Alternatively, an intervention can be "soft" in which case it supplies the variable intervened on, *X*, with an exogenous source of variation that is not correlated with other causes of *X* (those causes that are not on any route from *I* to *X* to *Y*) but does not break the connection between *X* and those other causes (cf. Eberhardt and Scheines, 2007). Note that the interventionist condition (**M**) does not say that causal relations are present only when interventions are actually performed. Rather it connects the existence of causal relationships to what would happen *if* interventions *were* to be performed. From this perspective, when one reasons with non-experimental data to causal conclusions, one is trying to use the data to predict what would happen if certain experiments were to be performed but without doing experiments. I will say more about this below.

My emphasis in what follows, however, will be not so much on the role of interventions per se but rather on certain other ideas intimately associated with interventionism—particularly on various notions of *independence* and *invariance*. I will attempt to show how these notions

---

[2] For stylistic reasons I will sometimes flout mathematical precision by using "correlated" to mean "statistically dependent".

[3] But with the following difference that will be important in subsequent discussion: My view is that an intervention *I* on *X* with respect to Y must be implementable by some $I \rightarrow X$ relationship that is "distinct" from the $X \rightarrow Y$ relationship. When this is not the case, an intervention on *X* with respect to *Y* is not possible. As I understand him Pearl does not impose such a condition. See Section 5 below.

connect both to the asymmetric features of causal relations and to interventionist treatments of causal claims. In doing so I hope to cast light both on the asymmetries and on the significance of invariance/independence notions for understanding causation. I stress that what follows is *not* intended as an argument for an interventionist account of causation. Rather, I going to assume that something in the neighborhood of this account is correct and then use it to try to illuminate some features of explanatory and causal asymmetries.

2. Some Preliminaries

To motivate and explain this project, I begin with the observation that in one sense the asymmetries under discussion can be captured or represented perfectly well just by linking claims about causal direction to claims about what happens under interventions. Suppose that *X* and *Y* are statistically dependent and assume that if *X* (*Y*) causes *Y* (*X*), *Y* (*X*) does not cause *X* (*Y*)[4]. (Here "intervention" can understood either as "hard", or when this is more appropriate (see below) as "soft".) Then if (i) *X* causes *Y* there will be an intervention on *X* that changes *Y* and no intervention on *Y* that changes *X*[5] while if (ii) *Y* causes *X* holds, there will be an intervention on *Y* that changes *X*, while no intervention on *X* will change *Y*. Applying this idea to the flagpole case, we can argue as follows: in the case of the flagpole, *H* and *A* cause and explain *S* because (i) intervening on *H* (e.g., by shortening the height of the pole) will change *S*, (ii) intervening on *A* (e.g., by replacing the sun with a different light source at different angle to the ground) will change *S* and (iii) by contrast, intervening on *H* will not change *A* (showing that *H* does not cause *A*) and intervening on *S* (perhaps by putting up a barrier which prevents illumination of the pole by the sun) will not change either *H* or *A*, showing that *S* does not cause either of these variables.

Alternatively (and to anticipate discussion below) we might reason in terms of "soft interventions" as follows: Suppose that we confine ourselves to the example as originally discussed by Hempel and others, and thus assume that *H*, *A* and *S* are the only relevant variables (there are no omitted common causes of *H, S* and *A* and that the goal is to capture the difference between the following two alternatives: (i) *A* and *H* cause *S* or (ii) *A* and *S* cause *H*. We may then reason that if, in accord with (i), the causal direction is from *H* and *A* to S, *A* will be a soft intervention variable on *S* in circumstances in which *H* and *A* are statistically independent (since *H* is constant for any given pole, this condition will be satisfied as long as *A* varies, which will happen over the course of the day). Of course under such interventions on *S* via *A*, we observe no changes in *H*. Assuming that (i) and (ii) are the only alternatives and given the other

---

[4] At least at the level of type causation, I think that systems in which *X* causes *Y* and *Y* causes *X* are entirely possible—see Woodward, forthcoming for examples. But for purposes of this paper I assume that we are not dealing with systems for which this is the case.

[5] "No intervention on *Y* that changes *X*" is meant to cover two possibilities: it may be (i) that it is possible to intervene on *Y*, but under any such intervention there is no change in *X*. Alternatively, it may be impossible to intervene on *Y*—that is, there is no way of satisfying the conditions for an intervention on *X*. See below for additional discussion.

assumptions above there is no other candidate for a variable that might be used to intervene on *S*, so we infer that *S* does not cause *H* and hence that (i) is correct.

Treatments of this sort seems correct as far as they go, both as accounts of what the differences between (i) and (ii) imply about what would happen if various interventions were to be performed and also as accounts of how, by performing such experiments, we could conclusively establish what the correct explanatory direction is in the flagpole example. Nonetheless they are less than fully satisfying. For one thing, both in the flagpole example and in a number of others discussed below, we seem able to reach correct conclusions about causal and explanatory direction without performing interventions, relying just on observational (non-experimental) information about, for example, dependence or correlational relations of various sorts among variables, perhaps in conjunction with other sorts of assumptions. (Details of how this might work and what other sorts of assumptions are needed will be presented shortly.) This suggests that there are features – call them F-- present in such examples that we use to correctly infer causal and explanatory direction even if we have not performed the appropriate experimental interventions. As a matter of epistemology and methodology it is important to understand what these features F are and how they figure in inferences regarding causal direction. As we shall see, this is a very active area of research in statistics and machine learning, among other disciplines.

A second consideration which reinforces the first is this: the notion of an intervention is of course itself a causal notion and as such has a notion of causal direction built into it—the causal direction goes from the intervention *I* to the variable intervened on. For this reason if someone is puzzled about the notion of causal direction itself, appeals to what would happen under interventions as a way of understanding causal direction will seem less than fully satisfying[6]. (When I speak of puzzlement about the notion of causal direction this includes, for example, questions about whether causal direction is "objective", having its source in how matters stand in the world or whether instead it is in some way reflective of facts about us and our "pragmatic" interests, as suggested by philosophers as different as Hempel, 1965 and Price, e.g., 2018.) One way of addressing this puzzlement is to attempt to connect the directional features of causal claims to other important features that causal relationships possess. I will take these to include the aforementioned features F which guide inferences about causal direction in non-experimental contexts. As I suggest below, we can think of these features F as (or as connected to) structural features in the world that "support" or provide bases for claims about causal direction. In other words, my claim is that understanding the bases on which we make inferences about causal direction can help us to better understand some of puzzling features of causal direction itself. In what follows, I will argue, following similar ideas in the machine learning literature, that these features F have to do with various notions of invariance and independence conditions which many causal and explanatory claims satisfy.

Before doing this, however, a methodological digression is in order. Some writers who have discussed causal direction frame their discussion around a contrast between, on the one hand, an underlying "metaphysics" having to do with what causal direction "is" or what it

---

[6] For example, Dowe, 2019, who writes that because of its non-reductive character, "interventionism doesn't tell us what the direction of causation is" (45)

"consists in" and, on the other hand, mere "heuristics" that may be epistemically or methodologically useful for inferring causal direction but which have no bearing on what causal direction is, metaphysically speaking[7]. (It is often supposed that an answer the "is" question requires a reduction of some kind.) I don't think of the ideas that follow as a contribution to the metaphysics of causal direction, although that no doubt depends on what one understands as metaphysics. I do, however, claim that the invariance/ independence features F can be understood as contributing to what might be described as aspects of the worldly infrastructure that supports claims about causal direction[8]. As I see it, the procedures for inferring causal direction that I will describe "work" because they pick up on and extract information about the independence/invariance features F associated with causal direction. In this sense these procedures are not "mere heuristics" or superficial cues that are only of epistemological significance. In other words, *how* we find about causal direction is intimately connected to *what* we find out about when we find out about causal direction.

In suggesting this I am trying to point (admittedly unclearly) to a third possibility besides the metaphysical project of specifying what causal direction is and the project of providing mere heuristics which are at best relevant to the epistemology of causal direction. This is the project, alluded to earlier, of elucidating the worldly infrastructure that underlies and grounds assessments of causal direction. I see this project as connecting epistemological concerns having to do with how we find out about causal direction with the "what is out there" concerns of metaphysicians, although my answer to the what is out there question does not involve any kind of elaborate metaphysics. My general picture is that causal thinking "works" to the extent that it does because it picks up on or is supported by certain generic features of our world, including in the case of the directional aspects of causal thinking, the features F alluded to above.

I will add that my view is that the supporting features in question are ordinary empirical features which, although often present in our world, will not hold in all logically possible worlds and are not usefully thought of as reflecting conceptual truths. One consequence is that my discussion of causal direction is *not* intended to apply to worlds that are wildly different from our own: For example, I will not attempt to capture "intuitions" some may have about what

---

[7] I am grateful to Marc Lange for pushing this point of view in a characteristically clear and courteous way in correspondence. Marc's assessment is that the ideas discussed in this paper are relevant to the epistemology of causation but not to its metaphysics. I agree, at least on some conceptions of what metaphysics involves but, as I go on to say, I think there is another possible project, besides metaphysics and epistemology to which I hope to contribute.

[8] If you want to regard this as metaphysics, that is fine with me. I will add that to my ear, talk of what causation or causal direction "consists in" or what these "are" or what "constitutes" them sets up the expectation that there is some "material" or "stuff" out of which these are "composed". Such questions about constitution make sense in many cases (e.g., one can sensibly ask what gold consists of) but my view is that causation and causal direction are not like this. Instead we need to understand them functionally. We can, however, sensibly talk about the worldly structures that support or allow us to make sense of causal direction and this is what I attempt to do.

causal direction amounts to in universes that contain just two particles. To the extent that a metaphysics of causal direction attempts to address questions about what causal direction consist of in all possible worlds this is not my project.

Having said this, I also want to insist that, independently of what one thinks about the infrastructure project, the epistemological/methodological problem of how one finds out about causal direction in contexts in which experimental manipulation is not possible is an interesting and important one in its own right—both from a philosophy of science perspective and because of its connection with many other disciplines interested in causal inference.

Two further points: First, I suggested above that when one infers causal direction on the basis of non-experimental information what one is in effect doing is inferring what would happen if various interventions were to be performed without actually doing the interventions, relying instead on other features present in such situations – the independence/invariance features F. We should thus think of the features F not as an alternative to the interventionist account of causal direction but rather part of the same package. My basic test for causal direction is the interventionist one described above. I see the features F as relevant to causal direction because they can furnish information relevant to questions about what would happen under interventions. More subtly (as I will try to elucidate) these features help to underwrite the very possibility of interventions.

Second, let emphasize that the relationship between causal and explanatory direction and the invariance/ independence features F I will be exploring is *not* proposed as a way of "reducing" the directional features of causal and explanatory claims to invariance/independence claims. For one thing we require a notion of causal direction to properly state the invariance/independence claims. Rather my goal is to "make sense" of the directional features of causal or explanatory claims (or at least some of them) by relating them to various other features possessed by causal claims—additional worldly structure associated with such claims.

Given this conception of the project several other consequences follow. First, I see no reason to suppose – and so will not argue—that there is some single source of the directional features of causation. The treatment that follows accordingly discusses several distinct, albeit related considerations that are relevant to causal direction. Moreover, I do not claim that these are the only features that are relevant to causal direction—there are others that I do not discuss[9]. Second although the independence/invariance features on which I focus are satisfied by many scientific theories or causal analyses they are not satisfied by all successful theories. For example, one source for causal directionality has to with independence assumptions among initial conditions. But some forms of this assumption such as assumptions about the independent assignability of initial conditions everywhere along a Cauchy surface of the sort contemplated by Wigner (discussed in sections 4-5) will not be satisfied by theories like classical electromagnetism and general relativity that are not purely hyperbolic in form and contain constraint equations. I do not regard this as problematic for my account. My view is that when

---

[9] For example, I do not discuss directional features that are present when a more general theory explains another as a special case. Here there is typically an asymmetry in derivability relationships.

certain independence features are satisfied, we can appeal to them to illuminate causal direction. When these features are not present, then, if there is well-defined causal direction, it must be understood in some other way. Since the project is to describe connections and worldly supporting structures, we are not required to find universal necessary and sufficient conditions for causal and explanatory directionality.

Finally , the examples I discuss in this paper are macroscopic—flagpoles, samples of gas and so on. Some writers suggest that the directional features of causation are present only in macroscopic systems and are not to be found in microscopic systems. For the most part little will turn in this paper on whether this claim is correct. I'd count it as a success if what I say about causal direction works for macroscopic examples (which I insist are interesting and important in their own right). But that said, I see no reason to suppose that the independence/invariance assumptions to which I appeal and the treatment of causal direction which follows from them holds only for microscopic systems. For example, independence constraints on initial conditions can certainly hold for systems involving atoms and molecules. In general, the idea that we can only make sense of causal direction at a macroscopic scale seems very implausible. When beams of protons collide within one another (*C*) in the LHC and various scattering events and products occur (*E*) does anyone doubt that the causal direction runs from *C* to *E*[10]?

The rest of this essay is organized as follows. In Sections 3-4 I briefly discuss and put aside two alternative suggestions about causal asymmetries. The first is that these have the source in "pragmatic" considerations. The second is the asymmetries can be fully understood in terms of time order. Sections 5 and 6 introduce two independence/invariance conditions that are closely bound up with causal direction: value/ relationship independence (**VRI**) and statistical independence of causally independent initial conditions (**CSI**). Sections 7 and 8 apply **CSI** to several familiar examples including the flagpole case. Section 9 explores some relationships between **CSI** and strategies from the machine learning literature for inferring causal direction in additive error models. Section 10 discusses some examples illustrating the relationship between value/relationship independence and causal direction. Section 11 draws some general morals from the previous discussion about how the directional features of causation sometimes arises, locating this in the *relationship* between initial and boundary conditions and governing laws, rather in the latter taken alone. Section 12 extends the framework developed in previous sections to asymmetries in non-causal explanations.

3. Pragmatics.

A number of authors[11], including Hempel himself, have treated the directional features of causation and explanation as a matter of "pragmatics". Exactly what this means is far from straightforward (and no doubt varies from author to author) but in the present context I take the idea to be that the directional features we ascribe to explanations and causal claims have their

---

[10] For a critical discussion of the claim that asymmetries governing causal direction apply only at the macroscopic level, see Frisch (2014).

[11] See also van Fraassen, 1980.

source in facts about human psychology (perhaps in facts about our "interests" or what we chose to focus on). Or, relatedly, perhaps the directional features derive from a particular "perspective" that we adopt as temporally located deliberating agents (Price, 2018). Or perhaps they are rooted in highly contextual features of the systems under analysis of a sort that elude more systematic specification. In any case the intended contrast is with more "objective" features, specifiable in a systematic way and independently of facts about human psychology. This contrast is reflected, for example, in the way that Hempel introduces the notion of "pragmatic aspects" of explanation (1965, 425). These are taken to vary depending on the characteristics of the persons involved in the process of explaining—with what they happen to find intelligible, illuminating or relevant (1965, 426) -- in contrast to more "objective" features of explanations that do not exhibit this sort of relativity to persons. Hempel's view is that these objective features don't provide a basis for judging that explanations of effects in terms of their causes are superior to explanations of causes in terms of their effects— this is what he has in mind when he describes the directional features as a matter of pragmatics. We may find the cause → effect explanations more satisfying or natural than effect -> cause explanations[12] but if so but this is just a fact about human psychology or perhaps just a fact about the psychology of some of us.

My view is that the best response to this challenge is to identify features that are "objective" and that distinguish causes and effects and explanations of effects in terms of their causes from those that work in the opposite direction. In other words, I see Hempel and a number of other philosophers who have advocated "pragmatic" treatments of causal and explanatory directionality as arguing by default; they think that there are no objective grounds for such judgments of directionality (or at least none that elucidate how directional features contribute to some objectively characterized notion of explanatory goodness) and hence opt for a pragmatic treatment in the absence of any other alternative. One can thus show that the pragmatic treatments are unnecessary or unmotivated by providing the kind of objective account that Hempel and others think does not exist—this is what I aim to do. Of course one of the best ways of arguing for the "objectivity" of causal directionality is to show that there are procedures that reliably identify causal direction and that make use of information about how matters stand in the world, rather than information about our interests or about human psychology[13]. And one of the best ways of arguing for the claim that there is an objectively correct notion of (causal) explanatory direction is to show that getting causal direction right has explanatory significance.

---

[12] I'm eliding some distinctions here. One might think that the directional features of causation are objective but that they have no explanatory significance. This may have been Hempel's view: causal directionality is grounded in time order but explanations of causes in terms of their effects can be just as good as the reverse.

[13] This provides one illustration of my claim that considerations having to do with how we find out about causal direction can have implications about how we should understand causal direction—that is, that the former are not of "merely epistemological" significance. In other words, if there are strategies for successfully identifying causal direction that work by picking up on such objective features as, e.g., patterns of correlation, then this argues for the objectivity of causal directionality.

I will add that even those who find pragmatic accounts initially attractive ought to find objective accounts of causal direction of value if they can be shown to exist.

4. Time order.

Another common suggestion about the direction of causation/ explanation takes this to be fully grounded in time order considerations: According to this position, if the only two alternatives are that (i) *X* causes *Y* or that (ii) *Y* causes *X*, (i) will be true if *X* or instances of *X* temporally precede instances of *Y* and (ii) will be true if the temporal order is the opposite. For example, it might be argued (cf. Salmon, 1984) that because of the finite speed of the propagation of light, the shadow cast by a pole will come into existence a short time after the light source that produces the shadow is switched on and this is why the direction of causation/explanation is from the former to the latter.

It is certainly true that in many cases we make (and are justified in making) judgments about causal order based on time order considerations[14]. But as a general account of causal direction, the appeal to time order is unsatisfying for several reasons. First, there are many cases (some discussed below) in which we make judgments about the direction of causal explanation in the absence of time order information, which suggests that we must be relying on other sources of information in making such judgments. In some of these cases, there may be "underlying" facts about temporal order but either we do not know these or do not seem to rely on them in making judgments of causal direction. In still other cases, the variables with which we are working may not be defined in such a way that we can order them temporally, so that there are conceptual barriers to using time order to sort out causal direction[15]. These considerations are reflected in the fact that the problem of inferring causal direction without relying on information about temporal order is recognized as a major problem in many disciplines, including machine learning and econometrics. The procedures for inferring causal direction described below do not rely on time order.

An even more fundamental problem is that such accounts provide no insight into (or justification for) *why* time order should matter in the way that it does in explanation and causal judgment. Consider, for example, Hempel's view of the flagpole problem. He is perfectly aware

---

[14] In addition to other illustrations, time order information can be used in combination with other inference principles such as the Causal Markov condition (discussed briefly below) and a causal minimality condition to infer causal relations—in some cases, permitting identification of a unique set of such relations. See, for example, Pearl, 1988, Hitchcock 2018, Stern, Forthcoming. (The minimality condition requires that when a graphical model M satisfies the Causal Markov condition with respect to a probability distribution P, no proper submodel of M satisfies the Causal Markov condition with respect to P.)

[15] This may happen if, for example, the variables do not have sufficiently fine-grained temporal locations to distinguish competing claims about temporal order. This may be true, for example, for variables defined over extended temporal intervals—e.g., GDP per quarter. In other cases variables may not be temporally indexed at all, as is the case with variables measuring personality traits in social psychology.

that some DN derivations are such the *explanans* variables take their values before the *explanandum* variable takes its value, while others have the opposite profile. He asks, in effect, why this should make any difference to the explanatory status of the derivations. In fact, it clearly shouldn't if, as Hempel, thinks, explanation is just a matter of deriving an explanandum from laws and other conditions. A satisfactory response to Hempel needs to show what getting the directional features right contributes to correct explanation and causal judgment. Appeal to time order as a primitive basis for sorting out causal or explanatory direction does not do this. Put differently, what we are looking for is (i) an account of causal explanation and causal claims – an account of what such explanations *do* when they are good and (ii) an associated account of causal direction that enables us to understand what (ii) contributes to (i). Skeptics about "objective" treatments of explanatory direction such as Hempel haven't been answered until we have done this[16].

This is also the appropriate place to correct a misunderstanding about the relationship between time order considerations and interventionist interpretations of causation and directed graphs. The notion of an intervention *I* on a variable *X* presupposes, as I have said, a notion of causal direction: the causal direction is from *I* to *X*. However, the notion of an intervention of *I* on *X* does not build in (at least in any obvious way) assumptions about time order[17]. That is, as far as the technical notion (taken in itself) of an intervention *I* on *X* goes, *I* might be temporally located after *X* or (more plausibly) there may be no well-defined temporal relation between *I* and *X*. This is reflected in the fact that there is no reference to time order in standard characterizations of the notion of an intervention. Relatedly when one claims that some relationship is invariant under interventions one is not building in a reference to time order. Similarly, when one uses a directed graph to represent a causal relation between *X* and *Y* (*X*→*Y*), and gives this an interventionist interpretation, this means that the direction of causation is from *X* to *Y* but it does not (or at least we need not take it as implying) that *X* temporally precedes (or is not later than) *Y*. Of course we think that in most, perhaps all cases, causes do not occur after their effects but this idea is not built into interventionism[18].

---

[16] Even if you are tempted to say that it is true by some definition of causation that effects cannot precede their causes, there is still the question of why we operate with a notion of causation that has this feature. Why shouldn't we replace our current notion with some notion that permits backward causation or that is undirected? In other words, what work (if any) does the idea that causal relations have a distinctive direction do for us? Saying that we call the event that comes first the cause does not explain the significance of causal direction.

[17] Here I disagree with Ismael, 2016.

[18] Some may think that this is a defect in interventionism but I think it is a virtue. For one thing, there are physical theories that are often interpreted as claiming that causes occur after their effects (The Wheeler -Feynman absorber theory and the Lorentz-Dirac equation of motion for charged particles in classical relativistic electrodynamics are commonly mentioned candidates— See Earman, 1976.) Such theories may not describe our world but it is not obvious that they are conceptually incoherent. An account that builds into *X* causes *Y* the requirement that *X* cannot occur later than *Y* judges such theories to be obviously incoherent, so that they can be immediately rejected on apriori grounds. My contrary inclination is to think that if backwards

Since the focus of this essay is on considerations relevant to causal direction that are not based on time order considerations, there are many interesting and important questions relating time and causation that I do not address, at least in any detail. For example, there is the issue, noted immediately above, of why our world apparently does not contain instances of "backward" causation in which effects temporally precede their causes. There is also the general issue of the relation between casual directionality and thermodynamic asymmetries, including the connection of these with various cosmological hypotheses, such as the past hypothesis. I touch on this only very briefly in Section 13. My failure to discuss these issues in any depth does not mean that I regard them as unimportant. It is, however, also interesting that there is much that can be said about causal direction without directly discussing time and entropy.

5. Some varieties of Independence and Invariance: Value/ Relationship Independence

I turn now to a discussion of several different varieties of independence which I claim can be connected to causal and explanatory direction in illuminating ways. I distinguish three of these—(i) independence in the sense of statistical independence of variables that are causally independent (causal to statistical independence or **CSI**), (ii) independence between the values of cause variables and the causal relations/laws in which they figure (variable relationship independence/invariance or **VRI)** and, closely related to (ii), (iii) in dependence of different causal relationships from one another. My main focus will be (i) and (ii).

I begin with (ii) since this is the most natural point of entry. A basic feature of many physical theories and also of structural equation models that purport to represent causal relationships is a distinction or "cut" between what are often called "initial conditions" (hereafter ics) – "accidental" facts about the values certain variables happen to take --- and the laws or causal generalizations (hereafter c-generalizations) connecting variables, including those having to do with initial conditions, to one another.

Before proceeding two caveats are in order: First, I use "initial conditions" because this is common parlance; this usage is not meant to imply anything about temporal order. I might have instead described these as conditions represented by "independent" variables or by variables that represent causes, as opposed to effects[19]. Second, talk of "initial conditions" is not meant to deny that there are other conditions, including boundary conditions and constraints that are also important in constructing causal analyses and explanations, particularly when these involve differential equations. I will say a bit more about this below.

---

causation is incoherent, this will so for subtler reasons. Second, and relatedly, one would like to have a non-trivial explanation of why causes rarely if ever occur after their effects. Building time order into causal direction makes this impossible—the only possible "explanation" is that this is analytic given what we mean by "cause". There is a lot more to be said about this topic but not here.

[19] Think also of an initial value problem in the theory of differential equations where "initial" need not be understood in terms of temporal order.

In many cases it has proved possible to separate such initial conditions from the c-generalizations in such a way that they satisfy the following condition: the c-generalizations continue to hold—they are stable or robust—under various changes in the ics. In such cases I will say that the c-generalizations are *invariant* under changes in the ics. For example, initial conditions for application of the Newtonian gravitational law include the values of the masses *m1* and *m2*, and the distance *d* between them. The law itself continues to hold—that is, it continues to accurately describe what will happen—under changes in the values of these initial conditions, both those that occur in a single system and across different systems. Similarly for other sorts of changes—spatial translations and Galilean transformations of gravitating systems. Plausibly these invariance features are at least part of the reason why we regard the gravitational generalization as a law[20].

In the case of structural equation modeling it is standardly assumed that if an equation – e.g., $Z= aX+bY$—describes a causal (or genuinely "structural") relationship (with *X* and *Y* causing Z in the way the equation indicates), then this equation will continue to hold, under changes in the values of *X* and *Y* (think of these as corresponding to initial conditions) for at least some range of changes in these values. Of course equations meeting this condition in the contexts in which causal modeling techniques are used will typically hold under a much smaller range of changes in initial conditions than the generalizations we regard as physical laws but some degree of invariance of the sort described is plausibly regarded as a necessary condition for those equations to represent causal relationships.

Figuring out how to make the cut between initial conditions and c-generalizations such that the latter are at least to some extent invariant over the former is an extremely important step in constructing an explanatory theory in many cases. That we are sometimes able to separate c-generalizations and ics in this way and that the result allows for accurate predictions of the behavior of many systems is, as emphasized by Wigner, 1970 and others, a highly non-trivial fact and one that should not be taken for granted[21].

To make a connection with what will come later, another way of thinking about the invariance property just described is that involves a kind of *independence* of c-generalizations from initial conditions: the cut between c-generalizations and initial conditions is made in such a way that (ideally) they are independent of each other. "Independence" in this context obviously cannot mean statistical or probabilistic independence—c-generalizations are not random variables characterized by joint probability distributions involving initial conditions. Nor does it seem right to think of this sort of independence as a kind of causal independence, at least in any straightforward sense. As noted earlier, one way of expressing the basic idea is in terms of counterfactuals: the initial conditions should be such that they can change "independently" of

---

[20] See Woodward 2018b, 2020 for more detailed defenses of this claim.

[21] Wigner (1970): "The surprising discovery of Newton's age is just the clear separation of laws of nature on the one hand and initial conditions on the other". On the other hand, Wigner also makes it clear that he thinks it quite possible that this separation may fail in some (e.g., cosmological contexts)

the c-generalizations in the sense that the latter would remain the same (would continue to hold) were the former to change in various ways.

To make this more precise consider the contrast between the following two structures:

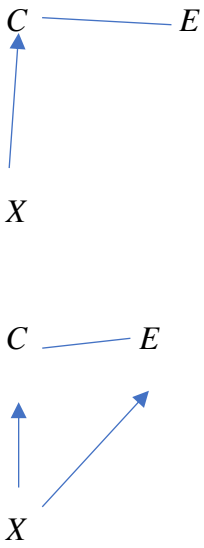


Figure 1

Directed arrows represent causal relations in both structures. In both structures there is a correlation between *C* and *E*, represented by the undirected edge. Suppose that in structure (i), *X* is the only cause of *C* and is uncorrelated with *E*. Then if *E* changes under changes in the value of *C* (where these are caused by changes in the value of *X*), this provides good reason to conclude that the correlation between *C* and *E* is causal. One basis for this reasoning is that in (i) the change in *C* due to *X* is intervention -like and the conclusion that *C* causes *E* follows from **M**. By contrast, if *E* changes under observed changes in *C* under (ii) this does not provide good reason to conclude that *C* causes *E*, since the correlation between *C* and *E* may be entirely due to the common cause *X*.

When we talk about the relation between *C* and *E* being invariant/independent under changes in the value of *C*, we should require that this invariance holds under changes in *C* that are caused in the way represented by (i) and not just in the way represented by (ii). This suggests:

> A necessary and sufficient condition for a c- generalization relating *C* to *E* to be invariant/independent of some range of changes in *C* is that the generalization would continue to hold if values of *C* were generated by causes *X* of *C* which are interventions with respect to *E*.

Suppose that we are able to determine that changes in initial conditions have occurred due to some appropriately intervention-like process like (i) and that we observe that some c-generalization continues to hold across these changes. This of would establish that the kind of independence/invariance under discussion is present. Suppose, by contrast, we observe a change

in initial conditions and that some candidate generalization continues to hold across those changes but we are not able to observe or directly determine whether those changes in initial conditions are the result of some intervention-like processes. That is, we observe a correlation between *C* and *E* under change in the value of *C* but not whether the changes in the value of *C* are caused by some *X* that has the properties in structure (i) or alternatively by some *X* in structure (ii). For example, we observe the joint probability distribution of two variables *C* and *E* (and that they are correlated) but don't observe the factors that determine *P(C)*. Given some candidate function *f* (c-generalization) linking *C* to *E*, is there some way of determining whether *f* is (in the sense under discussion) "independent" of *P(C)*? And if so, can we use this information to infer causal direction? Indeed, what might "independence" mean in this sort of case?

One way of approaching this, problem, employed in portions of the machine learning and (in a sense) in the econometrics literature, is in terms of the requirement that there be a kind of *informational independence* between the c-generalization and the associated initial conditions: information about the values of the initial conditions should not tell us anything specific about the c-generalization linking *C* to *E* and conversely. On my interpretation[22], this informational independence is treated as a kind of (perhaps fallible) surrogate for or indicator of the counterfactual notion of invariance/independence described above. Informally, we can think of informational independence as implying that there should be no specific constraint relations between the initial conditions and the associated c-generalization —an idea that can be made more precise in terms of algorithmic information theory, as noted below[23]. Wigner alludes to something like this idea in his 1970, when he writes that, ideally, there should be "no relation" between initial conditions and associated laws. Obviously some such absence of a constraint relation is implied when we require that laws or c-generalizations be freely combinable with different initial conditions.

We can connect this idea about absence of constraints between initial conditions and cp-generalizations to an explicitly interventionist treatment of causation in the following way: Suppose we are given a candidate c-generalization $C \rightarrow E$ and that it turns out that interventions that change the value of *C* are accompanied by associated changes in *E*. What this implies is that there is a way of generating values of *C* (a relationship R1 that allows for the causation of values of *C* from some cause of *C* such as *X* in (i) above) that is distinct or separate from the relationship R2 linking *C* to *E*. If there were no such relationship R1 that might be used to produce values of *C* where R1 is distinct from the $C \rightarrow E$ relationship, it would not be possible to

---

[22] That is, this is my attempt to elucidate some of reasoning underlying the techniques in questions, rather than anything that is explicitly said in this literature.

[23] As noted in Zhang et al. 2015 this absence of constraints idea is also closely related to various notions of "exogeneity" found in the econometrics literature. All of these represent attempts to capture versions of the idea that the relationship or process that generates the cause should be appropriately separate from or independent of from the relation connecting the cause to the effect. Information about the former relationship can then be used to provide information about the latter relationship.

intervene (in the technical sense) on *C* with the observed result[24]. To the extent that R1 and R2 are distinct relationships, this at least suggests that there should be no specific constraint relations between them. Metaphorically, we might think of this in terms of the idea that nature chooses c-generalizations and initial conditions via separate, independent processes which are not "correlated" with or" tuned to" one another. As we shall see, when a requirement like this is satisfied, it can sometimes allow us to make inferences about causal direction.

6. Statistical Independence of Causally Independent Initial Conditions.

So far we have been talking about "independence" of c-generalizations from initial conditions or causes. There are, however, additional independence conditions that sometimes seem very natural and that can be imposed on the initial conditions/causes themselves (once we have separated them out from the c-generalizations as described above). One such condition connects causal independence and statistical independence (**CSI,** as referred to earlier): suppose there are distinct random variables[25] *X1*.. *Xn* such that none of the variables in this set are causes of other variables in the set and none of these variables share common causes (i.e., they are causally independent or exogenous). Satisfaction of **CSI** requires that these variables be statistically independent[26]. Or to take the contrapositive, if we do find statistical dependence among these

---

[24] What about the converse? Suppose that we have evidence that there is a relationship R1 that might be used to cause values of *C* via some *X*, and that R1 is "independent" of the $C \rightarrow E$ relationship. By itself this is consistent with *X* being a common cause of *C* and *E* in accord with scenario (ii) above. Suppose, however, that we think that whenever a common cause structure is present, it must be possible in principle to interfere with the two joint effects independently of each other—that is, we can break any arrow from *C* to *E* while leaving the arrow from *X* to *C* intact. (This is a consequence of a commonly accepted requirement in the causal modeling literature, called Independent Fixability in Woodward, 2015, discussed also in footnote 27.) It follows that it will be possible to use *X* to intervene on *C*.

[25] In other words, we assume that the variables can be treated as though they conform to some probability distribution that allows us to make sense of claims about statistical independence and dependence regarding them.

[26] Note that this doesn't mean that "coordinated" behavior among independent causes on particular occasions is impossible; rather it means that its probability of this occurring is low, for the same reason that a long run of heads in a series of causally and statistically independent coin flips is possible but unlikely. A coherent wave converging on a point formed as the result of waves from a large number of causally independent sources is not impossible, but it follows from CSI that the operation of such sources will be statistically independent so that such convergence will be rare.

Let me add, since there seems to be some confusion about how such one-off cases of coordinated behavior should be understood, that I do not understand them as involving (or as evidence for) backwards causation or reversal of temporal direction or anything like that. When a

variables, then there should be a casual explanation for this, either in terms of cause/ effect relations among the variables themselves or in terms of common causes.

This is one version of what is sometimes called the principle of the common cause. Something like this is sometimes described in the physics literature as the assumption that "incoming" influences should be uncorrelated (if we understand incoming influences to be causally independent[27]). It is also endorsed or implicitly assumed in many uses of causal reasoning in social science. It is a consequence of (but strictly weaker than) the Causal Markov condition[28] that is widely assumed in the causal modeling literature. For example, in the case of equation like (5.1) $Z= aX+bY$ one assumes that either the two cause variables X (which are represented as causally independent as far as this equation goes) and Y are statistically independent or, if they are not, that there is some additional causal relationship (or relationships) not represented by (5.1) that accounts for this dependence— e.g., either X causes Y or conversely or they have a common cause. The representation of this additional relationship will require additional equations—the relationship is not represented by (5.1) itself.

Let me repeat that my claim is that CSI describes a generic pattern that, as a contingent empirical matter holds widely, if not universally, in our world. I do *not* claim that CSI reflects a conceptual or metaphysical truth of some kind that holds in "all possible worlds". My assumption is that CSI and similar principles, although contingent, help to underpin the ways in which we think about causation and causal direction. I will not speculate about how if at

---

coherent wave forms from independent sources, the causation involved is ordinary forward in time causation running from the sources to the wave front. I also do not hold (see Section 11) that such cases have an "equivalent description" in which the causal directions are reversed, so that the wave in question can equally well be described as incoming and as outgoing and caused by some event at the point of convergence.

[27] For a number of examples illustrating applications of this idea in physics contexts, see Frisch (2014). Let me add that "incoming" influences are often understood to be temporally earlier than their effects. Philosophers who deny that time has an objective direction are often led by this consideration to the conclusion that it is arbitrary (or involves a "double standard") to hold that incoming influences are uncorrelated while outgoing influences (assumed to occur later) are correlated. Whatever one thinks of this contention, it is important to understand that CSI is a claim about causal order, not temporal order. As subsequent discussion will make clear, the bases of causal order are at least somewhat independent of the *bases* of temporal order. As nearly as I can see, **CSI** is not undercut by claims about the unreality of temporal direction.

[28] A graph G and associated probability distribution P satisfy the Causal Markov condition (**CMC**) if every variable is probabilistically independent of its non-descendants conditional on its parents. This is much stronger than **CS1** since unlike **CSI**, **CMC** connects causal claims to conditional independence claims—common causes screen off their joint effects from one another etc.

all one think about causal direction in worlds which **CS1** is systematically violated (or which we might find it tempting to describe in that way)[29].

Note that **CSI** does not, as formulated, embody a temporal asymmetry. It connects causal and statistical independence but says nothing about causes occurring temporally before their effects or about independence being present before causes interact to produce an effect but not after[30]. Also **CSI** describes a sufficient condition for statistical independence but not a necessary one. In fact it is obviously possible, even common, for causes that have interacted in the past to be statistically independent or effectively so—this can happen if for example they also have lots of interactions with other, uncorrelated causes, so that correlations produced by the earlier interaction wash out[31].

Two further points: First, I will understand CSI as having, so to speak, an architectural or strategic component. Given a set of variables and associated causal relations for which CSI appears to fail, it will often be a good strategy to look for new variables and causal relations formulated in terms of them for which CSI holds. (I take this to be one of the themes of Wigner's discussion: we should try to *discover* initial conditions which are such that CSI or some similar initial condition holds.) Second, as already suggested, I assume that whether it is possible to do this is in a way that results in an empirically adequate theory is an empirical matter, which depends on what the world is like. It is not a conceptual truth or metaphysical necessity that it will always be possible to formulate successful theories or analyses satisfying CSI.

I will not try to defend **CSI** here—there is a big literature about this—but will assume that, whatever its limitations may be, it is applicable (leads to reasonable inferences) in an interesting range of cases. (That is, it *works*, whatever its ultimate justification and limitations may be.) One of my goals in this paper is rather to show how for systems for which **CSI** holds we can use this principle to reason about causal direction.

As noted above, the architectural aspect of **CSI** suggests that we should look for models or explanations in which the assumed initial conditions or the variables that are represented as exogenous are statistically independent of each other. One motivation for this is the thought that if such statistical independence among initial/exogenous conditions is not present, this is (according to **CSI**) an indication that our model is not complete; there must be further unrepresented causal relations that account for the dependence. Postulating dependencies among initial conditions without a causal story of how these arise is thus to be regarded as unsatisfactory

---

[29] Of course if the way in which we think about causation is not applicable to such cases, it presumably doesn't make literal sense to describe them in terms of violations of CS1 which does embody the way in which we think about causation.

[30] It does say that causes of effects have a different statistical signature than effects of causes but this involves a causal, not a temporal asymmetry.

[31] This important point is noted in Myrvold, 2020.

or at least as indicating unfinished business. By contrast a model in which there is independence of initial conditions represents a natural stopping place in explanation or causal analysis[32].

Although neither of the two independence conditions **VRI** and **CSI** make reference to time both require, for their correct statement, a notion of causal direction. In the case of **VRI,** the requirement is that the c-generalizations $C \rightarrow E$ linking cause to effect should be invariant under changes in the values of the *cause* variable C This is very different from (indeed, as we shall see, in many cases inconsistent with) the requirement that the $C \rightarrow E$ generalization be invariant under changes in the value of the effect E. In many cases this latter invariance claim is false.

A similar point holds for **CSI.** This requires statistical independence among *cause* variables (in the absence of causal relations connecting those variables) but of course it does not require statistical independence among effect variables. Given a structure that looks like this

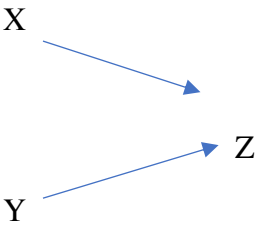


Figure 2

we expect, in accord with CSI, X and Y to be statistically independent in the absence of further information. On the other hand, if were to reverse the arrows to yield the following structure

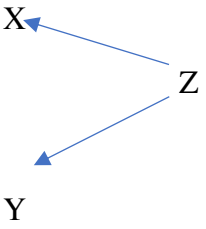


Figure 3

we would expect X and Y to be dependent.

---

[32] Wigner expresses this as follows—the initial conditions themselves "should be as random, as the externally imposed gross constraints will allow with the existence of regularities in initial conditions being considered "unsatisfactory" . (p 41)

It may seem tempting to infer from these observations that in order to use **VRI** and **CSI** we must have already identified the correct causal direction. In fact exactly the opposite is true --the features just described often make it possible to *infer* causal direction: Suppose, for example, we find that a candidate generalization relating *C* to *E* is invariant under changes in *C* (*C* → *E* is "independent" of the value of *C*)—something that, as noted above, can sometimes be determined empirically-- but (*E* →*C*) is not invariant under changes in *E*, Then at least in many cases we can conclude that the causal direction is from *C* to *E*. (See Sections 7-9) Similarly, given a case in which there are three variables, two of which are pairwise correlated and one pair of which is independent (as in Figure 3 above) , we can, given additional assumptions (see **P** immediately below), use **CSI** to infer that the direction of causation is from the two independent variables to the third.

7. Some Applications of CSI

I turn now to more explicit application of these ideas connecting independence to causal asymmetries beginning with the flagpole problem and **CSI.** Here I will make use of the following principle (which I take to be motivated by **CSI**):

> (**P**)[33] Suppose there are 3 variables, *X*, *Y* and *Z* such that either (i) *X* and *Y* cause *Z* or (ii) *X* and Z cause *Y*. (In other words there are no omitted common causes etc.) Suppose the patterns of dependence among these three variables are as follows: X_|_Y, X_/|_Z,

---

[33] This principle and the applications that follow are heavily influenced by Hausman, 1998 which remains one of the best discussions of causal asymmetry that I know. Hausman describes the "central intuition" of his account as the claim that "causal priority consists in the causal connection among the effects of a common cause and the causal independence of the causes of a given effect" (55). Two events are "causally connected" if one causes the other or they share a common cause; causal independence is the absence of causal connection. The similarity between this idea and **P** should be obvious. Nonetheless, there are differences. Hausman holds that causally connected events are "typically" statistically dependent and that causally independent events are not but that this is not always the case—on his account, the connection between causal independence and statistical independence involves a separate "operationalizing assumption". P does not have this feature. (The most obvious way in which statistical (in)dependence and causal (in)dependence can come apart involves failures of Faithfulness—see Spirtes et al. 2000.)_ Second, and perhaps more importantly Hausman's account is a proposal about what causal priority "consists in". I understand this to mean that if, for example, an event *E* does not have two independent causes, there is no fact of the matter about causal direction involving *E*. Principle P does not have this implication since it does not describe a necessary condition for there to be a fact of the matter about causal direction. In fact, as explained later, my view is that the causal direction can sometimes be identified when *C* is the only cause of *E*. Within an interventionist framework as long as there are possible interventions on *C* that will change *E*, *C* causes *E*; it does not matter if there are no other causes of *E*. Even when interventions are not performed and *E* has no other causes besides *C* causal direction can still be well-defined. On the other hand, as suggested above, this is not to say this notion is well-defined in all possible circumstances.

Y_/|_Z, where _|_ means statistical independence and _/|_ means statistical dependence. Then (i) is the correct causal order.

To apply this principle to the flagpole example, I will follow standard presentations of the problem in assuming that the only two alternatives are that *H* and *A* cause (or causally explain) *S* or that *S* and *A* cause *H*, so that principle (**P**) applies. (This conforms to the standard formulation of the problem which asks why we should distinguish (and prefer qua explanation) a derivation in which *H* is in the explanans from a derivation in which *S* is in the explanans.) Suppose that we observe several flagpoles of different fixed heights $h_1..h_n$, at different times of day for each pole, so that *A* varies. In this case for any given *A*, there will be a correlation between the heights of the poles and the corresponding shadows of lengths $s_1..s_n$ but no correlation between *H* and *A* . As *A* varies over the course of the day, we also find, for each pole, a correlation between *A* and the length of the shadow cast by that pole. Thus we have the following pattern of independence and dependence relations: *H_|_A , H_/|_S, A_/|_S*. Applying **P**, we infer that *H* and *A* cause *S*.

There are a number of different ways of thinking about the justification for (**P**) and its applicability to this case: First and most obviously, the above pattern of dependencies is what we should expect if

(i)H and *A* cause *S*

is the correct structure but not if

(ii) *S* and *A* cause *H*

is correct. According to (i) (and assuming that the alternative possibilities are restricted in the way described above) *H* and *A* are causally independent and hence by **CSI**, we expect *H_|_ A*. By contrast if (ii) is the correct structure then again by **CSI** we should expect *S _|_A*, which we do not observe.

Note that although this reasoning relies on **CSI,** it does not rely on anything stronger such as the Causal Markov condition or on the assumption of faithfulness **F** which is sometimes assumed in causal modeling[34]. In particular in connection with **F** we do not require the assumption that if *X* causes *Y*, *X* and *Y* are (statistically) dependent (which is an implication of **F**), but rather only a "converse" assumption according to which causal independence implies statistical independence. What enables us to avoid relying on a faithfulness-like assumption is that we have assumed that the only two alternatives are (i) and (ii). If we do not make this assumption and instead consider a broader range of possible alternative structures for relations among *H, S* and *A* then a faithfulness like assumption would be required to reach a reliable conclusion about causal direction. However, it is hard to fit many of these alternative structures

[34] A distribution is faithful to a graph if the only independence relations in the distribution are those that follow from the application of the Causal Markov Condition to the graph.

into any standard understanding of the flagpole problem, which is why a restriction to (i) and (ii) seems appropriate.[35]

We can also connect principle **P** with standard interventionist thinking and thus get further insight into why **P** "works" as follows. As noted above, within the interventionist framework the claim that *H* causes *S* and *S* does not cause *H* corresponds to the claim that there are interventions on *H* that will change *S* but no interventions on *S* that will change *H*. The claim that *S* causes *H* has the opposite profile concerning the results of interventions. Again assuming that these are the only two possibilities (and making the assumptions about the absence of common causes etc. described above), the pattern of (in)dependencies *A_|_H, H_//_S, A_//_S* suggests that *A* functions as a soft intervention variable on *S* , since it is exogenous and independent of the only other possible cause of *S*, namely *H*. Observation shows that changes in this intervention variable *A* for *S* are not associated with changes in *H*. This suggests that *S* does not cause *H*. Moreover, if we assume that *S* causes *H*, then, under this assumption, there will not be, among the variables in the system, any intervention variable for *H* that is independent of *S*, since the only remaining variable, *A*, is correlated with *S*[36]. In short the pattern of dependencies

---

[35] Gebharter (2013) shows how by applying the SGS algorithm (Spirtes et al., 2000) and assuming the Causal Markov and Faithfulness conditions one can derive the correct causal structure for the flagpole problem from the observed independencies. Gebharter's derivation is entirely correct. However, as noted, if we restrict the possible graphs to (i) and (ii) above, we don't need the assumption of faithfulness. More generally, assuming that our model is restricted to the three variables *H*, *A* and *S* , a violation of (triangle) faithfulness would arise in structures in which there is an arrow from *H* to *S*, and arrow from *S* to *A* and a cancelling arrow from *H* to *A* or alternatively and arrow from *A* to *S,* and arrow from *S* to *H* and a cancelling arrow from *A* to *H*. If, as is generally assumed in discussions of the flagpole problem, we know the functional relation among *H*, *A* and *S* (*S= H cot A)*, the only issue being identifying causal direction, these sorts of faithfulness violating cancelling structures are excluded. (In any case, no one thinks that it would make sense to suppose that, say, *H* by itself causes *S* and also by itself causes *A* via an independent route, with *A* in turn causing *S*.) That said, in more complex structures, faithfulness does real work in identifying causal direction and orienting arrows.

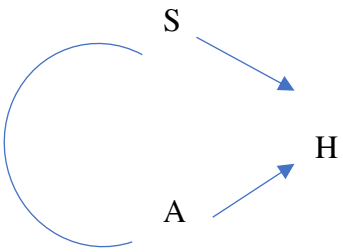
In this connection it is also worth noting an obvious trade-off: An advantage of using assumptions like Causal Markov and Faithfulness is that one does not need to restrict the hypothesis space in the way I have above. On the other hand, if we do restrict the hypothesis space we can get by with assumptions weaker than CMC and Faithfulness. I don't think that either strategy is necessarily better than the other—it depends on what you think that you know. In general, the machine learning strategies I discuss proceed in part by restricting the hypothesis space (e.g., by restrticing the functional forms considered or assuming the absence of confounding). This allows for results that would not be possible without such restictions.

[36] That is, if *S* causes *H* then one expects that it ought to be possible in principle to intervene on *H* by means of some intervention variable that is independent of *S*—this is an implication of a commonly assumed principle in causal modeling, called independent fixability (**IF**) in Woodward, 2015. **IF** says that it should be possible to intervene on every variable in a causal model, fixing its value independently of every other variable. Assuming that *S* causes *H*, there is

suggests that there is a route to changing *S* that is independent of *H* (which is what we expect if *H* causes S) – namely the route involving *A*-- but no route to changing *H* that is independent of *S*, which is what we expect if *S* causes *H*.

On this view of the matter, a pattern of (in)dependence relations involving *H*, *A* and *S* conveys information (given the background assumption that one is choosing among a very limited range of possibilities) about would happen if various interventions were to be performed, even though no interventions are in fact performed. This is an example of what I meant earlier in saying that (in)dependence information can be connected to interventionist ideas concerning causal direction in a way that illuminates how the former can be a source of information about the latter. It also illustrates how observational information, not involving interventions, can be used in conjunction with background assumptions to answer questions about what would happen if certain interventions were performed.

Another related way of thinking about the flagpole example, appeals to the desirability of avoiding unexplained coincidences or dependencies when there are equally adequate alternative models that do not require such coincidences. As noted above, when one observes a single flagpole, the naturally occurring changes in *A* over the course of the day due to changes in position of the sun will be correlated with *S*. Moreover, *S* and *A* will change in concert in just such a way that the value of *H* remains constant. Thus in a model in which *S* and *A* cause *H* (with no causal connection between *A* and *S*) *S* and *A* will appear to be precisely "tuned" to each other, varying so as to maintain a constant value for *H*, despite the absence of a causal connection between these variables. The model in which *S* and *A* cause *H* will thus look like Figure 4 with the double-headed arrow between *S* and *A* representing the fact that they co-vary together, despite the fact that neither is represented as causing the other and they are not represented as having a common cause.



---

no obvious candidate for such an *S*-independent intervention variable for *H*. It is true that **IF** requires only the possibility of intervention and it might be argued that this is consistent with the absence in fact of such a variable. However if the alternative possibilities are restricted to (i) and (ii) in the way described above, then there is no intervention variable of the required sort among the possibilities. This is at least suggestive that the assumption that *S* causes *H* is mistaken. Of course there are in fact (other) intervention variables for *H* and we may be able to observe them (or at least we may be aware of their existence). One most obvious candidate is the actions *X* of the person or machine who fashioned the pole as having one height rather than another. These will typically be exogenous with respect to the other variables under investigation. Such variables can also help with identifying causal direction as I note immediately below.

Figure 4

By contrast in a model in which *H* and *A* cause *S*, there is no such unexplained dependency: all of the observed dependencies follow just from the causal structure of the model and what are assumed to be exogenous changes in *A* (or in *H* if we are considering populations of poles.) In one obvious sense the model in which *S* and *A* cause *H* is less simple than a model in which *H* and *A* cause *S*—less simple in the sense that the former model requires additional information (in the form of a statistical dependency between *A* and *S*) besides the two causal arrows it postulates to account for the observed dependencies while the latter model requires only two causal arrows. There is thus a kind of redundancy in the $S \rightarrow H$ model since the observed dependencies could be accounted for without postulating the *A*—*S* correlation[37]. This theme—that models with the wrong causal direction typically involve additional unexplained coincidences or special "tuning" will recur in connection with other examples discussed below. It provides additional illustration of how accounts which get explanatory direction wrong seem deficient *qua* explanations[38].

There is another, related way of thinking about the flagpole example which will be useful later in our discussion. So far we have been considering causal and correlational relations just involving *H, A* and *S*. But (as noted in footnote 36) there is another source of information about causal direction. This has to do with variables that are exogenous causes of *H*. Often we have at least some information about these. (In realistic cases these often will be hard intervention variables for *H*.) An obvious candidate for such a variable is the actions/intentions *X* of the person or machine who fashioned the pole as having one height rather than another[39]. In some cases we may be able to observe such an *X* but even if we do not, we will often be confident about some of its characteristics, such as that it is an exogenous cause of *H*: in other words $X \rightarrow H$ and it is not the case that $A \rightarrow X$ or that $S \rightarrow X$. Since *H* and *S* are correlated, if we know that *X* is an intervention variable for *H*, this licenses the conclusion that *H* causes *S*. We can also reason in the following way: Suppose, for purposes of refutation, that *S* causes *H*. Then,

---

[37] Note that this is different from the kind of redundancy that is present when a model violates Minimality or Faithfulness. Also in this connection, Reuben Stern has drawn my attention to Forester et al. (2018). This argues for a criterion for model choice based on the idea that, ceteris paribus, models with fewer directed edges are preferable. The model with *H* and *A* causing *S* and the model with *H* and *S* causing <u>*A*</u> have the same number of directed edges but the latter has an additional undirected connection.

[38] The idea that it is a kind of defect in a model or theory if it involves special tuning or coincidences to capture observed results is common to many areas of science. However, there are many possible kinds of tuning and it is by no means obvious which are objectionable and why.

[39] What about the case in which the maker of the pole fashions it with the intention I that it cast a shadow of a certain length in a certain location at a certain time of day, as in van Fraassen, 1980? In this case it is *I* that functions as an exogenous cause of *H*. This is certainly not a case in which *S* causes *H* (or *X*).

assuming $X \rightarrow H$, H will have two causes S and X[40]. But then, unless (i) X causes S and S causes H, X and S should be independent and they are not. If one is seriously worried about possibility (i), it can be shown that the influence of X on H is not mediated by S by, for example, the observation that X has the same influence on H regardless of the value of S. That is, the influence of the person making the pole on its height is the same, regardless of whether a shadow is present. By contrast, conditional on H, X and S are independent which is consistent with the ordering being $X \rightarrow H \rightarrow S$[41].

As we see from this example, causal information about some variables, including information about causal direction can, when combined with correlational information, constrain causal direction among other variables. As we shall see in Section 12 a similar sort of strategy can work when in some cases involving non-causal explanatory dependencies.

8. A More Subtle Example

Now consider a more subtle example[42]. Suppose first (i) a sample of gas in a container with fixed volume V is placed in a heat bath of constant temperature $T = t$. Here the natural judgment is that V and T are causes of pressure P. Contrast this with following case: (ii) The gas is again placed in a heat bath at temperature t but the gas is now in a cylinder with a movable piston with surface area A. A weight W is placed on top of the piston. The gas is allowed to expand until it reaches an equilibrium at volume V in which the force F ($F = P.A$) due to pressure P is exactly balanced by the downward force of the weight W. Now the correct causal order seems to be that P and T cause V. Principle **P** gives the correct analysis of both examples. Again there are 3 variables which are causally related. In connection with (i) if we were to observe a "random" population of gas samples with different values of V and T (different fixed volumes and temperatures) we would see that V and T are uncorrelated but that both are correlated with P. If these are the only three relevant variables, we may infer in accordance with **P** that V and T are causes of P. In (ii) again looking at a random population of gas samples with movable pistons and variable weights , T and P will be uncorrelated (the pressure is causally fixed by W and the temperature by the heat bath which is causally independent of W) , while T and V and P and V

---

[40] There are other, more outre' possibilities such as X being a common cause of both H and S, with no causal relation between these last two variables. I will not explore this since we are assuming as background that either H causes S or S causes H.

[41] H being a common cause of X and S is also consistent with this conditional independence relation but I assume in typical cases that we can be confident that H does not cause X.

[42] An example having this sttucture is briefly discussed in Woodward, 2003 and in more detail in Hausman et al. (2013). These authors conclude, on the basis of the observation that causal direction seems to change in the system described below that the system "eludes causal representation" at least by a single directed graph. I agree that the causal relations in the system depend on what is held fixed and hence that no single directed graph describes the causal relations in the system across changes in what is held fixed. But I wouldn't describe this as a case in which the system eludes causal representation; rather different representations are appropriate, depending on what is held fixed.

are correlated so the correct direction is that *T* and *P* are causes of *V*. Note that the same "law" or c-generalization *PV=nRT* governs the gas in both cases (or so we can assume).

The two examples thus illustrate an important point that will receive more attention later. The causal direction in the examples is not just "in" (or fixed or determined by) the law *PV=nRT* considered by itself but rather (also) has to do with role played by the initial and boundary conditions and constraints governing the system. This includes information about what is or is not correlated with what among these conditions, but this in turn reflects what is fixed and not allowed to vary (as is the case with the container of fixed volume or the gas with the movable piston and fixed weight) and what is allowed to vary. That this information is relevant to causal direction is an implication of principle **P** since what quantities are correlated or not with others may depend (as the two gas examples illustrate) on what is fixed and what can vary in the specific systems we are considering[43]. As the example under discussion shows this information may not be contained just in the laws or c-generalizations governing the system, which is why different systems governed by the same law can exhibit causal relations with different directions. I will return to some of the implications of this observation below. Here I will just note that it would be a mistake to infer from this point that there is something unreal or non-objective about the causal direction present in these systems. The facts about what is correlated with what in different systems are indeed system-specific and "contingent" (in the sense of not being fixed by the laws) but that does not make them unreal or non-objective and does not make the associated claims about causal direction non-objective. "Objective" does not have to mean "fixed by the laws alone".

One way of thinking about the upshot of my discussion so far is that there is more content or structure present in many explanations and causal claims than what is captured by a simple focus on deductive relationships (or facts about "instantiation" of regularities) of the sort that characterize the DN model (and a number of other models of explanation). Information about which variables are independent of others (including, crucially, information about independence relations among candidate cause variables and which variables are to be regarded as fixed in value) contributes importantly to directionality and to explanatory import—this information is a "working part" of the explanation. Relationships that may look completely symmetrical (such as the relationship between the height of flagpole and the length of its shadow) can be shown to embody asymmetries when one attends to independence relationships. These asymmetries matter for successful explanation—they are tied to the ability of explanations to answer questions about what would happen if initial conditions were different (called w-questions in Woodward, 2003) and to the explanatory virtue of avoiding unexplained coincidences.

8. Causal Direction in Additive error models.

In the examples discussed so far, the causal relations are assumed to be deterministic and the values of all three variables figuring in those relations are observed. A body of recent work in machine learning (e.g., Janzig et al. 2012, Peters et. al. 2017, Shimizu et al., 2006, Hoyer et al,

---

[43] Similarly in the flagpole example, the fact the pole is rigid and of fixed height provides important information about causal direction.

2014) explores a set of different but related problems. Suppose that *X* and *Y* are statistically dependent but their relationship is stochastic or noisy where this can be represented by the presence of a noise or error term—i.e., *X* and *Y* are related by some function in which a noise term figures. We wish to determine whether *X* causes *Y* or conversely. We assume further that no unmeasured common causes are present and that the noise term enters additively into the relationship between *X* and *Y* , so that there are just two hypotheses about causal direction—either (i) $Y = f(X) + U$ or (ii) $X = g(Y) + U'$ where *U* and *U'* are error terms. We can observe *X* and *Y* but not *U* or *U'*. In one kind of case, the functions *f* and *g* are assumed to be linear but the processes that generate the candidate cause variables and the noise term are assumed to be non-Gaussian (more precisely at most one of these is Gaussian). A technique known as independent components analysis (ICA) which separates non-Gaussian distributions into statistically independent components is used to examine whether it is possible to fit an equation of form (i) to the *X*, *Y* distribution with $X \perp U$ and similarly to determine whether it is possible to fit an equation of form (ii) with $Y \perp U'$. If the error term can be made independent of the candidate independent or cause variable in one direction, but not the other, one infers that the former is the correct causal direction. The assumption of non-Gaussianity is crucial to the success of this procedure since ICA requires this assumption and more generally because the linear Gaussian case is symmetric—in this case it is always possible to fit independent errors in both directions so that the procedure gives no recommendations about causal direction[44].

In a second kind of case it is again assumed that no unmeasured common causes are present and that the relationship between *X* and *Y* involves an additive noise term, so that as before the alternative hypotheses are (i) $Y = f(X) + U$ or (ii) $X = g(Y) + U'$. However now the functions *f* and *g* are assumed to be non-linear. In this case if one can fit a model of form (i) such that $X \perp U$, then "usually" (with certain exceptions again including the case in which the joint distribution of *X* and *Y* is bivariate Gaussian) there is no such additive noise model in the opposite direction from *Y* to *X*—that is, no *U'* such that (ii) $X = g(Y) + U'$ with $Y \perp U'$. ("Usually" means that if (i) holds, the space of functions in which (ii) also holds is of much lower dimension.) Again if there is a model of form (i) with $X \perp U$ and no model of form (ii) with $Y \perp U'$ one infers that (i) is the correct model[45].

---

[44] One way of thinking about this is that a Gaussian distribution of a single variable contains relatively little information—the entire distribution can be characterized in terms of its mean and variance. Similarly for a bivariate Gaussian distribution (two means and variances and covariance information). In the case of a non-Gaussian distribution, information about higher moments is needed to characterize the distribution. This additional information, not present in the Gaussian case, can be relevant to causal direction. Ironically, given the tendency of main stream statistics to focus, until recently, on Gaussian distributions, non- Gaussianity actually aids causal inference.

[45] The authors suggest a way of making this general idea more operational as follows: First test whether a model of form (i) is consistent with the data by doing a nonlinear regression of *Y* on *X* , getting an estimate f^* for *f*, and using this to calculate the resulting residuals $U^* = Y - f^*(X)$, and then test whether U^* is independent of *X* . Then repeat the procedure with the model

Both of the methods just described have been tested on real world data for which causal direction is independently known (or at least there are generally accepted beliefs about this). Without going into a lot of detail, as an empirical matter, the methods perform reasonably well on many data sets, with accuracies in the neighborhood of 70 to 80 percent (as opposed to the 50 percent that would be expected from random guessing.) For example, given information about the correlation between altitude and rainfall in various areas in Germany, the first method correctly infers that it is more plausible that altitude causes rainfall than conversely. Given data on the duration of an eruption and the time interval between subsequent eruptions of the Old Faithful geyser in Yellowstone National Park, the method involving non-linear functions infers that the correct model is that "current duration causes next interval length" rather than conversely. (Note there is no reliance here on time-order information).

At an abstract level these methods closely resemble the methods described above in connection with the flagpole and gas cases. Both methods make use of statistical (in)dependence information with the guiding idea being that if there is independence among putative causes in one direction and no such independence in the other direction, then the correct direction is one in which the causes are independent. For example, when we find an error *U* which is independent of *X* but no error *U'* which is independent of *Y*, we infer that *U* and *X* are causes of *Y*. (Of course the additive error models also make use of additional assumptions, concerning the form of the function linking cause and effect as well as the distribution of the noise term, but in other respects they start with less information than in the previous examples—the error term is unobserved and must be inferred while all three variables are observed in the flagpole and gas cases. In effect the unobservability of the error term is offset by the additional assumptions made in the additive error model case.)

We can provide the same general diagnoses of why the machine learning techniques involving additive error models work that we appealed to in the previous examples. **CSI** suggests that causes should be independent in the absence of causal relations among them or omitted common causes. So if, e.g., *X* and *U* are independent and *Y* and *U'* are dependent, we take *X* and *U* to be causes of *Y*. In addition, the same considerations having to do with unexplained correlations apply: In a model in which *Y* and *U'* are claimed to cause *X* with *U'* and *Y* dependent there is an unexplained correlation between *U'* and *Y*. By contrast in a model in which *X* and *U* cause *Y* with *X* _|_ *U* there is no such unexplained correlation. Other things being equal, this favors the latter model.

Similarly, looking at the matter from an interventionist perspective, if, as we are assuming, the only two possibilities are that *X* and some *U* causes *Y* or that *Y* and some *U'* cause *X*, the existence of a *U* which is independent of *X* but not independent of *Y* strongly suggests that one can intervene on *Y* (by using *U*) without changing *X*, which is diagnostic of the absence of a causal relationship from *Y* to *X*. At the same time, assuming that there is some causal relationship R that determines the value of *X*, the independence of *U* from *X* in a relationship of form *Y*= *X*+

(ii). If the residuals are independent in one direction and not in the other, then one concludes that the correct causal direction is the one in which independence holds.

U also suggests that *R* does not affect *U*. This in turn suggests that these generating conditions R for *X* operate so as to change the value of *X* in a way that is independent of the other causes of *Y*, represented by *U*. Since if such changes occur, *X* and *Y* remain correlated, we have evidence that *X* causes *Y*. In other words, finding an independent error in on direction but not in the other amounts to finding relevant (soft) intervention variables, even if these are not initially observed[46].

10. Value/ Relationship Independence and Causal Direction

So far we have been considering cases in which the effect variable is the result of two[47] cause variables, where these may either be explicitly specified and observed (as in the flagpole case) or, alternatively, one variable may take the form of an unobserved noise term which is discovered through ICA or some other procedure. Remarkably, one can also sometimes determine causal direction even when there are only two variables -- a single candidate cause variable and a single candidate effect variable with no hidden noise term and even when the relationship between these is deterministic and, moreover, even when the function from cause to effect is invertible.

I will first try to provide some intuition regarding the basic idea and then describe some details. First recall the independence relation **VRI** discussed above, concerning the "independence" of initial conditions and the c-relationships relationships in which they figure. As noted above, "independence" in this context cannot mean statistical independence; instead in parts of the machine learning literature (e.g. Janzig et al., 2012) independence is instead understood as a kind of *informational* independence or more formally in terms of "algorithmic independence" defined in terms of Kolmogorov complexity. I will relegate details about the latter to a footnote[48] and here will stick with a more informal description and a motivating example. Suppose, as a specific illustration, that the causal relationship from *C* to *E* can be represented as a function *E= f(C)*. Then the idea behind **VRI** is that specific information about the distribution of values for *C*, the putative cause or explanans variable, which might be given in the form of, say, a probability distribution for *C*, or some generating function for *C*, should not provide information (at least of a non-generic sort) about the function *f* and conversely. When the causal direction is from *C* to *E*, this informational independence condition can be

---

[46] I am grateful to Kun Zhang one of the discoverers to the techniques in question, for helping to clarify this connection with interventionist thinking.

[47] A number of the results described can be extended to cases involving more than two putative cause variables but I will not pursue this, since I am interested in the underlying principle.

[48] The Kolmogorov complexity *K(s)* of a string *s* of bits is the shortest program that generates *s* using a previously specified universal Turing machine. The conditional Kolmogorov complexity *K(t/s)* of string *t* given string *s* is the length of the shortest program that can generate *t* from s. *t* and *s* are algorithmically independent if *K(t/s) = K(t)*. Let *s** be the shortest description of *s*. Then the algorithmic mutual information of the strings *s, t* is defined as *I(s:t) = K(t)-K (t/s*)*. Informational independence between the initial conditions and a c-generalization can then be understood as the mutual information between them being zero (up to some additive constant or small number).

shown, as a matter of mathematics, to "usually" fail in the other direction. That is, with f in $E=f(C)$ independent of information about C, g in $C = g(E)= f^{-1}(E)$ will usually fail to be independent of information about E—"usually" means that the set of functions for which such failure occurs has low dimensionality in some relevant space of possible functions. Informally this can be motivated in the following way: If the correct causal direction is from C to E, with $E=f(C)$, and the distribution of C and f are "independent", then "usually" both the distribution of E and the relation $g(E) = C$ will reflect the influence of both the distribution of C and the action of f on C. Metaphorically, we can think of f and the distribution of C as acting as a common cause or common influence on g and E, leading to a dependence between g and E. This suggests a heuristic according to which the correct causal direction for a set of (X, Y) pairs is the one for which the distribution one of the variables is "independent" of the function describing the relationship between this variable and the other variable while the incorrect direction is the one for which this independence condition does not hold.

To further illustrate the underlying idea, let me switch to a different example[49]: the context is now indeterministic and there are just two binary variables which are statistically dependent. The only possibilities are that X causes Y or that Y causes X. There are two possible factorizations

$$Pr(X, Y) = Pr(X)\, Pr(Y/X)$$

$$Pr(X,Y) = Pr(Y)\, Pr(X/Y)$$

In such a context it is natural to take the independence or invariance of the conditional probability $Pr(Y/X)$ under changes in $Pr(X)$ (where by changes in $Pr(X)$ I mean a change from one probability distribution $Pr_1(X)$ to a different distribution $Pr_2(X)$ – i.e., the distribution of X is not stationary.) as encoding information about the causal relationship, if any, from X to Y. That is, if the causal direction is $X \rightarrow Y$, then $Pr(X)$ should be independent of $Pr(Y/X)$ and $Pr(Y/X)$ should be invariant under changes in $Pr(X)$ and conversely. If instead the causal direction is $Y \rightarrow X$, then $Pr(X/Y)$ should be independent of $Pr(Y)$ and invariant under changes in this probability distribution.

It is relatively easy to see that invariance/independence in one of these directions under some specified set of changes in the cause variable is inconsistent with invariance in the other direction under the same set of changes given some very natural additional assumptions. Suppose that the conditional probability $Pr(Y/X)$ is invariant under changes in $Pr(X)$ and focus on the case in which X and Y have just two values, 0 and 1. Assume that $Pr(Y=1/X) \neq 0$ or 1 (for either value of X) and that $Pr(Y/X=1) \neq Pr(Y/X=0)$ which is plausibly a necessary condition for X to be causally relevant to Y. We have from Bayes' theorem

[49] This is my own example. It doesn't come from Janzig et al. and similar work. I claim it illustrates the same basic idea as their examples, but if this is wrong, it is my mistake, not theirs.

$$(10.1)\quad Pr(X=1/Y=1) = \frac{Pr(X=1)Pr(Y=1/X=1)}{Pr(X=1)\,Pr(Y=1/X=1) + Pr(X=0)\,Pr(Y=1/X=0)}$$

Suppose *Pr (Y=1)* changes in value. We want to know whether the conditional probability on the l.h.s of (10.1) will remain invariant under this change, given the assumptions that the probabilities *Pr (Y/X)* are invariant. Since *Pr (Y=1) = Pr (Y=1/X=1) Pr (X=1) + Pr((Y=1/ X=0) Pr (X=0)* and (we are assuming) the conditional probabilities *Pr (Y/X)* are invariant under changes in *Pr (X),* this change in *Pr(Y=1)* must involve a change in *Pr (X)*[50]. Since the conditional probabilities on the r.h. s of are assumed to be invariant, the value of the whole expression on the right must change, given the additional assumptions outlined above. Thus the value of *Pr (X=1/Y=1)* must change. A parallel argument holds for the other values of the conditional probability *Pr (X/Y)* under changes in *Pr(Y)*. Thus we see that if the conditional probabilities are invariant under a specified set of changes in one direction, they will not be invariant under those changes in the other direction[51].

---

[50] Recall that we are assuming that the conditional probabilities *Pr (Y=1/X=1)* and *Pr (Y=1/X=0)* are not equal and similarly for *Pr (Y=0/X=1)* and *Pr (Y=0/X=0)*. If these conditional probabilities are equal it follows that *X* and *Y* are independent, contrary to assumption.

[51] Here is another way of thinking about this example and the associated argument which was suggested to me by Jiji Zhang. Suppose one obseves a change in the joint distribution *Pr (X,Y)* and that one is willing to assume that this change is due to an intervention on one of these variables. Suppose also that it is observed that *Pr(Y/X)* is invariant under this change. This shows that the intervention was not on *Y*, since if it were, *Pr(Y/X)* would have changed. At the same time *Pr (X)* changes and from the argument above, we know that *Pr (X/Y)* is not invariant under this change. So we infer that the intervention was on *X* and that *Y* changes under this intervention, establishing that the causal directionis from *X* to *Y*.

Let me also add that what the argument in the text above shows is that if *Pr (Y/X)* is invariant under some change in *Pr(X),* then for the associated change in *Pr (Y)* implied by this change in *Pr (X), Pr (X/Y)* will not be invariant under this change in *Pr (Y)*. In other words, one and the same change to the joint distribution *Pr(X,Y)* cannot be a case in which *Pr(Y|X)* is invariant across the change in *Pr(X)* and also be a case in which *Pr(X|Y)* is invariant across the change in *Pr(Y)*. However, it remains possible that *Pr(X|Y)* is invariant under some changes in *Pr(Y)* and *Pr(Y|X)* is invariant under some *other* changes *Pr(X),* involving a different change in the joint distribution. If we are willing to also assume that there are just two possible alternatives—either (1) *Pr (Y/X)* is invariant under all changes within some range of values of *Pr (X)* or (2) *Pr (X/Y)* is invariant under the associated range of changes in Pr(Y), then the argument above establishes that only one of these alternative holds. Many thanks to Jiji Zhang for helpful correspondence regarding this point and for correcting a misinterpretation of mine.

Given the relationship between finding invariant relations and correctly identifying causal structure this helps to motivate the assumption that in this sort of case the correct causal direction is given by the direction in which the conditional probabilities are invariant. That is, in a two variable case meeting the conditions just described if *Pr (Y/X)* is invariant under changes in *Pr(X),* we should infer that the direction of causation is from *X* to *Y*. We thus see that, just as in the flagpole case, the fact that certain quantities are invariant or independent of other quantities can be used to establish asymmetries in what might otherwise look like symmetric situations.

This example also provides an illustration of what would be involved in initial conditions and a c-generalization being "tuned" to one another in such a way that **VRI** fails. If the causal direction in the example is $X \rightarrow Y$, then, as *Pr(Y)* changes, *Pr (X/Y)* will also change or adjust systematically in such a way that that the invariance of *Pr(Y/X)* under *Pr (X)* is preserved—changes in *Pr (Y)* will be tuned to changes in *Pr(X/Y)*.

In the case as just described we assumed that there was an actual change in the probability distributions *Pr (X), Pr (Y)* and considered which of the conditional probabilities were invariant under these changes. If we could observe such changes and the relevant conditional probabilities we could use this to infer causal direction. This strategy is employed by Hoover, 2001 in a series of papers investigating the causal direction between economic variables[52]. In other cases we may lack information about whether a change in the marginal distributions *Pr (X)*, *Pr (Y)* has occurred. All that we observe is the joint distribution at a given time. Nonetheless one might think that it still makes sense to ask about informational independence between *Pr (X)* and *Pr (Y/X)* and between *Pr (Y), Pr (X/Y)* and that if such independence holds in one direction but not the other, conclude that the former is the correct causal direction. One way motivating this is to reinterpret the argument above in informational terms: when *Pr (Y/X)* and *Pr (X)* can change independently of each other they will be informationally independent. In such cases we should expect that changes in *Pr (Y)* will be accompanied by changes in *Pr(X/Y)*—these two quantities will be "correlated" or informationally dependent or will seem "tuned" to each other. Finding informational independence between *Pr(Y/X)* and *Pr (X)* and dependence for *Pr (X/Y)* and *Pr (Y)* is thus a clue that the causal direction runs from *X* to *Y*.

I remarked above that in the machine learning literature, these ideas about informational independence can be represented in terms of algorithmic information theory. This allows for the formulation of a notion of informational independence in terms of Kolmogorov complexity that is analogous to statistical independence and that applies to objects that are not random variables (such as functions and probability distributions). Within this framework, with a candidate cause

---

[52] For example, it is observable that size of the money supply *M* and the price level *P* are correlated but it is a controversial question which, if either of these, causes the other. Hoover explored the behavior of the observed relation between money and prices under shifts in federal reserve policy concerning the money supply—he assumes that some of these shifts are intervention-like and amount to a change in the distribution of P (rather than just different draws from the same distribution). He argues that in such cases, if *M* causes *P*, one would expect that the relation between *M* and *P* would remain stable under shifts in *P*. He finds that this is not the case but does find evidence for causation in the opposite direction from *P* to *M*.

X and a function that *f* that generates *Y* from *X*, the independence notion can be stated as the requirement that the description of *X* should be algorithmically independent of *f* or perhaps algorithmically independent of *f* conditional on some specified body of background knowledge. Although this yields a way of formalizing informational independence and the proof of theorems about it, it is not helpful in the analysis of particular examples, since Kolmogorov complexity is not computable. Practical implementation requires a more operational notion of informational independence.

Here the literature (e.g. Janzig et al. 2012) appeals to more specific mathematical facts, relating various functional forms, including the following: Suppose that *X* and *Y* are real variables where $Y = f(X)$ is a differentiable bijective function on the [0, 1] interval with a differentiable inverse f^{-1}. If $\log f'$ and $P(x)$ (the probability density of *X*) are "independent" in the sense that

$$\int \log f'(x) Pr(x) dx = \int \log f'(x)\, dx$$

then $\int log\,(f^{-1})'$ and $Pr\,(y)$ are positively "correlated", i.e.,

$$\int \log\,(f^{-1})'\,(y) Pr(y)\, dy > \int \log\,(f^{-1})'\,(y)\, dy$$

unless *f* is the identity.

This suggests a test for directionality that consists in looking for "dependencies" between the derivative *f'* of *f* and the density of the candidate cause variable—in other words one looks at the relation between *f'* and $Pr(X)$ and between $(f^{-1})'$ and $Pr(Y)$. If, say, the former pair are informationally independent and the latter informationally dependent, one takes this as a reason to conclude that the correct causal direction is from *X* to *Y*. As an illustration (Janzig et al. 2012) suppose that *X* and *Y* are related as in Figure 5, with $Pr(X)$ uniform and $Pr(Y)$ highly non - uniform:

Figure 4: If the structure of the density of P_X is not correlated with the slope of f, then flat regions of f induce peaks of P_Y. The causal hypothesis $Y \to X$ is thus implausible because the causal mechanism f^{-1} appears to be adjusted to the "input" distribution P_Y.

Figure 5

Consider the regions of large slope for f^{-1} (small slope for f). These are "correlated" with large peaks for Y, as shown in the diagram. Given the uniform distribution of X, the regions in which f has small slope will transform values of X in those regions to very similar values of Y, so that the density of Y piles up around those values. In this sense there will be an informational dependence between f^{-1} and Pr(Y) – the slope of f^{-1} tracks the lumpiness of Pr(Y). By contrast, given the uniform distribution of X, there is no such "correlation" between Pr(X) and f. Thus one concludes that X causes Y rather than Y causing X. Note that in this case just two variables are involved, rather than three as previously. Moreover, the functional relation between them is deterministic and invertible.

This method, like those considered previously, can be tested experimentally on real world data in which the causal direction is known on independent grounds. The method again correctly identifies causal direction at a rate well above chance (accuracy rates in neighborhood of 75 % depending on details of implementation) – for example, for sets of observations of water levels at various locations along the Rhine (where it is agreed that upstream levels cause downstream levels rather than conversely.)

This particular operationalization of informational independence obviously requires that the functional relations between cause and effect meet various conditions—the functions must be bijective, differentiable with differentiable inverses etc. In other cases, we may have reason to believe that the functions relating cause and effect will not satisfy these particular conditions but it may be possible to find some alternative operationalization that draws on the same underlying idea about independence of the process that generates the cause from the process that generates the effect being a clue causal direction.

As I have interpreted this method, it attempts to infer what would happen to the function relating cause and effect—in particular, whether this would remain stable under changes in the distribution of the putative cause—from relations of informational independence or their absence that are observed within a single joint distribution, as illustrated in Figure 5 above. Clearly even if it is right that whether or not the relationship $X \rightarrow Y$ is stable under changes in the distribution of X is a reliable clue regarding causal direction, there is additional inductive risk in trying to infer such stability from informational independence relations in the way described, where we don't actually observe what happens under distributional changes in X but merely try to infer what would happen were such changes to occur from a single observed distribution of X.

In particular, one worry one might have about the example in figure 5 is that that there are, after all, functions and mechanisms that take relatively non-uniform distributions and produce uniform distributions as outputs—think of gambling devices such as roulette wheels. In such cases, the correct causal direction will be from non-uniform Y to uniform X rather than from uniform X to non-uniform Y, as the method under discussion recommends for the example in Figure 5. In fact, however, a closer look arguably supports the analysis provided above. In non-uniform to uniform cases involving gambling devices the operative dynamics or mechanisms will take any one of a very large range of distributions of initial conditions (e.g., in some treatments any probability density over the initial conditions that is absolutely continuous) into a uniform distribution. Thus what is going on in such cases is that the dynamics is (largely) independent of the initial conditions after all, so that the initial conditions are causes and the

distribution of outcomes the effect. In other words, we have information about a non-uniform input→ uniform output relation that is stable under changes in input which makes it clear what the correct causal direction is. This contrasts with the information that is available in Figure 5 where we see only a single non-uniform distribution which is associated with a uniform distribution, so that the choice is between a cause-effect function that takes a uniform distribution as input and produces a non-uniform output (as any function with a non-constant derivative will do) and an alternative function that takes a non-uniform distribution as input and exactly undoes the non-uniformity in such a way as to produce a uniform output. When this is the only available information, it is not so obvious that the former choice is unreasonable. It might be argued that functions that undo non-uniformity to produce uniformity are "unusual".

In any case, my concern here is not to argue for this particular implementation of informational independence but rather to stress the general idea that independence/invariance understood in terms of **VRI** between the distribution of a variable or its generating mechanism can contain important information about causal direction. Moreover, if my argument so far is correct, this is not merely a superficial symptom that happens to be associated with causal direction. It instead involves a deep structural feature present in causal relationships (or at least many of them): it is exactly when the $X \rightarrow Y$ relationship is invariant under changes in X and or independent of whatever is responsible for the generation of the distribution of X values that we can use manipulation of X and the $X \rightarrow Y$ relationship as a way of changing Y. In the remainder of this essay I want to examine some additional implications of this idea and of **CSI**.

11. Directional Features as Arising from the Relation between Initial and Boundary Conditions and Governing Generalizations: Against the Cause-in-Laws Picture.

One general moral that can be drawn from the discussion so far is that the directional features of causation are closely bound up with facts about the initial and boundary conditions of the systems we are analyzing and the way in which these are related to or interact with the c-generalizations governing those systems. Thus in many cases, the directional features are not to be found in the governing c -generalizations alone. We saw this in connection with the gas cylinder example, in which systems with different initial and boundary conditions had causal relations with different directions, despite being governed by the same law. Similarly **VRI** is obviously a condition concerning the relationship between initial conditions and candidate c-generalizations.

This general picture contrasts with a common alternative picture that that is explicitly or tacitly assumed by many philosophers. I call this the "cause in laws" picture. According to this picture, laws of nature (or more generally, governing c-generalizations, whether or not they are laws), taken by themselves, have rich causal content and directly describe causal relationships. Thus the "logical form" of such generalizations or laws is something like: "All *F*s *cause G*s", where "cause" has all its usual connotations, including directionality[53]. In other words, these generalizations themselves supply all the causal information (including information about causal

[53] It is arguable that the common expression "causal law" builds in this assumption.

direction) relevant to understanding the systems to which they apply, without any of this information coming from other sources. Explicit in endorsements of this position can be found in Davidson, 1967 and Armstrong , 1997. Moreover it appears to be implicitly assumed by the many other philosophers who write as though if causal notions have any legitimate role to play in science the generic features of such notions including directionality must be found or grounded in laws or c- generalizations alone.

It is well known that this picture generates a number of puzzles. First, the word "cause" or equivalent expressions does not explicitly occur in most fundamental physical laws—perhaps in none, depending on what one counts as a law. "Cause" also fails to occur in many c - generalizations employed in sciences outside of physics.

Another more fundamental problem concerns the apparent tension between the directionality or asymmetry of the causal relationships and various "symmetries" of most basic laws. "Symmetry" in this context is used in several different ways. Some writers use it to refer to the fact that fundamental laws are "deterministic" in both temporal directions: from past to future and from future to past. More commonly "symmetry" concerns the time reversal invariance of fundamental laws. (Which of course is different than bi-directional determinism.) Very briefly, characterization of time-reversal requires specification of an operation on the variables within an equation that replaces these with their temporal "inverses": the time variable t is replaced by $-t$, the velocity variable v by $-v$ and (according to most) in classical electromagnetism the magnetic field B should be replaced with $-B$. An equation or law L is then time reversal invariant if, when some physical process P is consistent with L, so is the time reverse of L. For example, according to the laws of classical electromagnetism, an accelerating charge will be associated with electromagnetic radiation radiating outward symmetrically from the charge. These laws also permit the time-reversed process according to which a spherically symmetric wave of electromagnetic radiation converges on a single charge which then accelerates —a process which appears to be rare, absent some special contrivances.

A number of philosophers have thought that time reversal invariance and other sorts of symmetries present in fundamental laws raise problems for the directional or asymmetric features of causal claims; the concern is that there appears to be nothing in fundamental physics that "grounds" or serves as a basis for these directional features.

This in turn has led to several different responses. One is that this shows that the assumption that causation has directional features is a mistake since there is nothing in reality that might serve as a basis for these features. Another possible response (perhaps not sharply distinct from the first) is that that since the directional features (allegedly) have no basis in fundamental physics, they must have some other source—one suggestion is that they derive in some way from facts about us such as a particular perspective we adopt as deliberators. Views of this are defended by Price, 2007, 2014 and are discussed by Ismael, 2016 among others.

A very different view of the status of the directional and perhaps other features characteristic of causation is that their apparent absence from fundamental physics shows that the equations of physics, in their usual formulation, require additional supplementation in the form of various free-standing "causality principles" that provides those equations with causal content. Such principles might be thought to be at work when, for example, certain solutions to an equation expressing a physical law are discarded on the grounds that they violate the

condition that effects cannot temporally precede their causes. Yet another possibility is to reinterpret the equations themselves so that they make straightforward causal claims —e.g., Coulomb's law may be interpreted as the claim that charges *cause* electromagnetic forces or fields that operate on other charges. Views of this are perhaps suggested in Cartwright (1983).

I think that all of these views rest on the mistaken adoption of the cause in laws idea. That is, advocates of these views assume that if a basis for causal notions (and in particular the directional features of causation) are to be found anywhere in science or in physics, they are to be found in physical laws (or perhaps other governing c-generalizations from sciences besides physics) alone. Not finding such a basis in laws, these writers look for the basis in more anthropocentric sources, or in causal supplements in addition to physical laws as ordinarily formulated or, alternatively, conclude instead that there is no basis. As explained above, my contrary suggestion is that the basis for the directional features of causation is to be found in facts about initial and boundary conditions characterizing the systems we are analyzing and how these relate to (or interact with) laws and cp- generalizations. At least some of these facts are captured by conditions like **VRI** and **CSI.** Arguably these conditions involve straightforwardly "objective" facts that describe how matters stand in the world—they are not somehow due to our human perspective or projective activities. At the same time, the idea that making sense of causation requires that free standing causal principles or additional causal interpretations be added to basic scientific laws is also unnecessary. Again, laws and governing generalization along with initial and boundary conditions, as ordinarily understood and without any need for supplementation are all that is required[54].

There is of course another strategy for attempting to make sense of various asymmetries we find in the world (entropic and otherwise, including causal asymmetries) This agrees that we need initial and boundary conditions (or at least what looks like these) as well as more familiar laws to generate the asymmetries. However this strategy appeals to a single boundary -like condition which is imposed just once on the early universe. This is the Past Hypothesis (e.g. Albert, 2000), according to which the very early universe was in a state of very low entropy. For reasons having to do both with space and my own competence, I will not discuss this strategy here. However, I do wish to note that it differs from the considerations to which **VRI** and **CSI** appeal. The latter appeal to facts about the "local" initial and boundary conditions characterizing specific typically small systems – flagpoles, gases in cylinders with pistons that may or may not be movable and so on, rather than to some global cosmological condition. This is not intended as a criticism of the past hypothesis but it does underscore that appealing to it is different from the considerations explored in this essay[55].

---

[54] That is, **CSI** and **VRI** involve ordinary characterizations of initial conditions and how these relate to c-generalizations—they are not add-ons that go beyond the physical facts characterizing those conditions. When, for example, initial conditions are causally or statistically independent, this is just an ordinary physical fact about those conditions.

[55] I take it to be true, as an empirical matter, that the universe began in a low entropy state and that this fact figures in an explanation of why the universe has various global features. This is different, however, from the claim that we need the past hypothesis to understand (at least in any

Another way of putting this general idea about where causation is "located" (or at least often located) is as follows: to the extent that laws and other governing generalizations are expressed in differential equations, causation is not "in" these equations taken alone but rather in the *solutions* to those equations which arise when we combine them with specific assumptions about initial and boundary conditions[56]. In particular, as emphasized by Earman (2011), the time-reversal invariant character of most of the fundamental equations of physics is consistent with particular solutions to those equations exhibiting various asymmetries, including asymmetries having to do with causal direction—indeed, studies of many such equations show that "most" of their solutions are asymmetric (Earman, 2011). The asymmetry in the solutions arises in the same way it does in the gas in cylinder example -- because of the way in which initial and boundary conditions we impose interact with the laws themselves to yield solutions that are asymmetric.

As an additional illustration consider again the contrast between the case in which diverging electromagnetic waves are emitted by an accelerating charge and a case in which a coherent spherically symmetric wave comes in from infinity and converges exactly on the charge. The difference between these two scenarios does not fall out of Maxwell's equations themselves but instead also has to do with the different initial and boundary conditions characterizing the two scenarios. In the diverging wave scenario, if the charge begins accelerating at t0, it is typically assumed that the relevant boundary conditions at infinity (or at some considerable distance from the charge) are that there is no electromagnetic radiation at t0 or at earlier times. In the converging wave scenario, by contrast, the boundary conditions involve a coherent wave converging on the charge at some time prior to t0. This asymmetry, combined with Maxwell's equations themselves, gives rise to the different causal judgments we make about the two scenarios—in the first, the accelerating charge causes the diverging wave, in the second the arrival of the converging wave causes the charge to accelerate.

Of course it is true that the scenario with the converging wave rarely occurs while the diverging wave produced by the accelerating cause is more common. As I see it, this reflects the sorts of considerations that underlie **CSI**—the idea that causal independence leads to statistical independence. Absent some special contrivance, production of a coherent incoming wave would

---

very direct way) the directional features that are present in the flagpole and other similar cases. I take my discussion to cast doubt on this claim

[56] I don't deny that causal direction may be built into to some laws, taken in themselves, without any contribution from initial conditions. For example, this may be true of F=ma. I just claim that there are a number of laws for which this is not true. It is also the case, as emphasized by Wallace (unpublished), that many generalizations governing the behavior of macroscopic systems, both in physics (e.g., the Langevin equation) and elsewhere are not time- symmetric, either in the sense of being time-reversal invariant or in the sense of being deterministic in both directions. I will not attempt (do not know how) to connect this last fact to issues about causal directionality, although I will note (as also observed by Wallace) that the derivations of such generalizations from underlying laws that are time symmetric typically involves assumptions about the absence of special tuning among initial conditions, assumptions that violate CSI and similar conditions.

require a very precise pattern of statistical dependence or coordination among causally independent sources and hence is very unlikely although not impossible. By contrast, when additional fields are absent, it is not surprising that distinct segments of the wave front of an outgoing wave are correlated because this can be traced to a common cause (the accelerating charge). It is for this reason that if we are given a snapshot of the charge as it begins to accelerate and another snapshot of the coherent wave at some distance from the charge and no information about which occurred first and asked to infer which of these is the cause and which the effect, we can confidently infer that the acceleration of the charge caused the wave rather than vice-versa. This reasoning is very similar to the other examples of reasoning about causal direction described earlier in this paper.

According to this interpretation, the diverging, outgoing wave scenario and the converging incoming wave scenario describe distinct physical processes. The physical basis for difference between the scenarios is not to be found in the law governing the scenarios which is the same for both but rather in the facts involving the different initial and boundary conditions that characterize the scenarios. Some writers (e.g., perhaps Price and if I have understood him correctly) claim on the contrary, that the two scenarios do not really correspond to different possibilities—the account in terms of the accelerating charge causing the outgoing wave and the account in terms of the converging wave causing the acceleration are just different, equivalent descriptions of the same situation[57]. The argument for this claim is that it is required by the time reversal invariance of the governing laws. TRI is interpreted as similar in status to a coordinate transformation so that the description in terms of incoming and outgoing waves just represent different representations on the same process. This is a problematic interpretation of TRI[58] but even putting that consideration aside, the argument just described again appears to assume that if there is any basis for features of the scenarios having to do with causal direction, they must be found in fundamental physical laws alone. We have rejected this assumption. Indeed, it seems to me that this assumption *must* be rejected if we are to make sense of the observed facts. In particular, it is a fact that the converging wave scenario occurs a lot less frequently than the outgoing wave scenario (and similarly for other scenarios requiring coordinated action by many independent causes—broken vases reassembling, gas molecules uniformly distributed throughout a container assembling in one corner and so on). If what we are dealing with is just two different descriptions of the same situation it not easy to make sense of this apparent difference in frequency of occurrence. On the other hand the difference in frequency of occurrence makes sense if we regard the two scenarios as genuinely different where this difference includes a difference in causal direction traceable to differences in initial conditions. Again, this is not to claim that causal direction is independent of the underlying physical facts since included among those are facts about initial and boundary conditions[59].

---

[57] If I understand Farr, he holds that there nonetheless is a single description that is most appropriate for characterizing both situations—this is the one in which accelerating charge causes the outgoing wave.

[58] See Earman, 1974.

[59] In other words, although TRI tells us that in many cases a process P and its time reverse P* are both possible, it does not imply that causes of P and the causes of P* are the same or that the

I suspect that one of the main reasons why the contribution of initial and boundary conditions to causal direction has been missed is that such conditions are widely thought by philosophers to modally inert and lacking anything relevant to causal content. Since causal claims, including claims about causal direction, presumably have modal content, it is natural to think that this content must be supplied entirely by laws or c-generalizations. The mistake in this reasoning is the assumption that facts about initial and boundary conditions and relations among these are modally inert. That this is a perhaps most obvious in connection with examples like the gas in a cylinder in which it is specified that the volume of the container *can* or *cannot* change. But it is also true that independence assumptions like **CSI** and **VRI** carry modal commitments. When it is assumed that different variables, used to specify the values of initial conditions, can change independently of one another, these claims have modal content. Similarly for claims about the independence of various generalizations across changes in initial conditions Thus *both* claims about initial and boundary conditions and how these relate to laws as well as the laws themselves carry modal commitments.

12. **Directionality in Non-Causal Explanations**

My discussion so far has focused on causal directionality and directionality in causal explanations. Recently there has been an upsurge of interest in non-causal explanations of various sorts. Let us assume, for the sake of argument that are such explanations or at least that this is a possibility worth taking seriously. Against this background, the question of whether such explanations have directional or asymmetric features and if so, how we should understand these, becomes important. One way of motivating this question is to note that, however in detail this is understood, causation clearly has directional features. But if an explanation is non-causal, then if it has directional features, these can't be causal in character. They must instead be understood in some other way. This in turn suggests an argument against the very possibility of non-causal explanation: Suppose that explanation of any kind must be asymmetric—if *X* explains (causally or non-causally) *Y*, then *Y* cannot also explain *X*[60]. In the case of causal explanation, we have a story about where this directionality comes from—it comes directly from the directionality of causation. But in the case of non-causal explanation we have (it is claimed) no similar story—no way of making sense of (no basis for) their directional features. Since explanation must have a

---

causes of one are the "time reverse" of the other or (at least in the macroscopic cases with which we are concerned) that both sets of causes occur with equal frequency. Indeed, TRI does not say anything about causation—it is not a transformation that acts on causal direction by "reversing" it (or failing to reverse it). This will seem particularly obvious if, as I have argued, causal direction is not essentially tied to time order. This assessment contrasts with Farr (2020) who asks whether the time reversal operation leaves causal direction invariant or not. He argues that the operation should be understood as leaving causal direction unchanged, so that a process and its time-reverse have exhibit the same causal relations. My view is that the causal relationships (including the directionality of such relations) are very different when a vase struck by a rock shatters into pieces and when the pieces reassemble into an intact vase that emits a rock—as I say above, this difference underlies the difference in frequency with which such processes occur.

[60] As noted earlier, I do not endorse this thesis as a general claim.

privileged direction, we should for these reasons reject the claim that there is such a thing as non-causal explanations[61].

One way of responding to this argument is to deny that explanation must (always) be asymmetric. However, a number of the most plausible examples of non-causal explanation in the literature do appear to have a distinctive direction (see below). Thus the issue of how if at all these directional features might be understood arises in a natural way—indeed an account of this seems to be required if we are to make sense of many of the supposed examples of non-causal explanation.

In this section I want to briefly explore the possibility of providing such an account by extending the claims developed in previous sections. My basic idea is that in a number of cases the directional features of non-causal explanations can be understood in terms of generalizations or extensions of the ideas about independence and its relation to directionality described previously. I will consider two examples—my treatment of them will be somewhat different but will share a common core.

One plausible candidate for a non-causal explanation is Euler's graph theoretical explanation of why it is impossible to traverse the bridges of Konigsberg via a continuous path in which each bridge is crossed exactly once (an Eulerian path). I will call this explanandum the transversability of the bridges, represented by a variable T that can take two values depending on whether or not the bridges are transversable. Since the Konigsberg example has been extensively discussed, I will assume that it is unnecessary to provide details. Suffice it to say that Euler identified a graph theoretical feature F which he proved to be necessary and sufficient for an Eulerian path to exist—the absence of this feature F implies that no Eulerian path exists and hence T has the value= non-transversable. The arrangement of bridges in Konigsberg does not possess the feature F. If we let E be a two valued variable representing whether feature F is present and assume for the sake of argument we are dealing with an explanation of some kind, one has the strong intuition that it is the graph theoretical feature E that explains T rather than vice-versa. In my (2018a) I argued that this directionality could be understood in terms of the following consideration: Although the explanation of T in terms of E is non-causal, there is a straightforward causal explanation for whether one value or another of E holds—this has to do with the intentions and behavior X of those who constructed the bridges[62]. In other words, $X \rightarrow E$ where the arrow here represents "causes". Now suppose that that T non-causally explains E (rather than E non-causally explaining T). It then would follow that E has two distinct explanations, one causal and the other non-causal. In my (2018) I stopped at this point, thinking that it should be obvious why this two explanation story ($X \rightarrow E$ <-- T (where the second dashed

---

[61] Something like this argument is advanced by Craver (2016). The discussion that follows can be thought of as a response, attempting to show how to make sense of the directional features present in at least some non-causal explanations.

[62] In other words although the relation between E and T is non-causal and one cannot intervene on E with respect to T, the relation between X and E is causal and one can intervene on X with respect to E.

arrow represents non-causal explanation) was less plausible than an account in which the direction of non-causal explanation runs from *E* to *T* (*X* → *E* --> *T*).

In a recent paper Lange (Forthcoming) criticizes this suggestion, claiming that there is nothing in the interventionist account that rules out the possibility that *X* causally explains *E* while *T* non-causally explains *E*. I agree with Lange that my argument rests on additional assumptions about how non-causal explanation work and how these interact with causal explanations. Let me try to make these explicit. I argued above that in a structure in which *X* is an intervention-like cause of *Y* (so that *X* and *Y* are statistically dependent), *Y* and *Z* are statistically dependent and *X* and *Z* are statistically dependent (where the intervention-like character of *X* is understood to rule out the possibility of confounding by additional common causes (no *W* that is a common cause of *Y* and *Z* etc.), it is reasonable to conclude that the causal direction runs from *Y* to *Z* rather than from *Z* to *Y*. The contrary conclusion – that *Z* causes *Y*—does not explain why *X* and *Z* are dependent and instead postulates two independent causes of *Y* that happen to be correlated with each other, but where no explanation is provided for this correlation. My suggestion is that in the absence of some specific reason to think otherwise, it is reasonable to assume that structures that involve both causal and non-causal explanations will obey a similar principle. That is, if, in the Konigsburg bridge example, *X* causes *E* and *E* and *T* and *X* and *T* are statistically dependent as they clearly are, then, *at least in the absence of some further explanation of these dependencies,* we should infer that the direction of non-causal explanation runs from *E* to *T* rather than conversely. (I will say more shortly about the qualification introduced by the italicized phrase.) The contrary assumption—that *E* has two explanations, one in terms of *X* that is causal and the other in terms of *T* that is non-causal but where *X* and *T* just happen to be correlated even though no explanation is provided for this fact -- is less plausible.

What about the italicized phrase above? This qualification is necessary because it seems possible that an explanandum *M* might have two explanations, one, E_1, that is causal and the other, E_2, that is non-causal[63]. The principle I propose does not deny this but rather claims that when this is the case and there is a systematic association or dependency between E_1 and E_2, there should be some explanation for why this is the case. For example, one possibility is that E_1 and E_2 involve characterizations of the same system but at different "levels" or scales, with the factors cited in E_2 supervening on or involving a coarse graining of the factors cited in E_1. (Think of statistical mechanical and thermodynamic explanations of the same explanandum.) In such a case, because of this supervening/coarse graining relation, there is no mystery about why there is a systematic relation between E_1 and E_2. My point is that the relation between *X* and *T* in the Konigsberg bridge example is not like this. *X* and *T* appear to be at the same "level", and neither is a coarse graining of the other[64]. There is no obvious reason why they should be associated in the way that they are if they are independent explanations.

---

[63] My argument is thus not that if *E* has a causal explanation, it follows automatically that it cannot have a non-causal explanation. I agree with Lange and other writers that causal and non-causal explanations of the same explanandum are possible.

[64] Consider the two explanations of the movement of a toy balloon in an accelerating airplane described in Salmon, 1989 and cited in Lange, forthcoming—one "causal", molecular and

This reasoning rests on the assumption that reasoning about directionality in non-causal explanation obeys, in the respect described, a similar principle to that employed in reasoning about causal directionality. Of course this assumption may be wrong but (i) it yields what most suppose to be the "right" answer in this case (as well as in a range of other cases of alleged non-causal explanation) and (ii) there is a rationale for the assumption when it is understood as an extension of a principle that applies to causal explanation. Someone who wishes to deny the assumption owes an account of non-causal explanation that shows why the assumption fails.

A second putative example of non-casual explanation, discussed far more tentatively in Woodward (2018) concerns the explanation of the stability (of perhaps the *possible* stability) of the planetary orbits in terms of the three dimensionality of space (in conjunction with assumptions about the form of the gravitational potential in spaces of different dimensions (that this involves a generalization of Poisson's equation) and Newton's laws of motion. Given the latter assumptions it can be shown that the orbits will be unstable in spaces of dimensionality greater than three, so that there is a sense in which the stability of the orbits appears to depend on the dimensionality of space. Woodard (2018) suggested that if one finds it plausible that this is an explanation (and thus that the correct direction doesn't run instead from the stability of the orbits to the dimensionality of space), this is likely because one is willing to make certain independence assumptions that parallel those that we make in the case of causal explanation. In particular one assumes that (i) Newton's laws of motion and the form for a generalized gravitational potential in an n-dimensional space are independent of (ii) the dimensionality of the space in the sense that (i) and (ii) can vary independently of each other. (This is the non-causal analog of the idea that the causes of an effect should be capable of varying independently of each other.) We appeal to this independence assumption when we argue, as envisioned in the explanation above, that if the dimensionality of space had been different from three, Newton's laws of motions and the form of the gravitational potential would have been the same. It is this assumption about independence, I claim, which allows us to give content to the contention that the correct direction of explanation runs from spatial dimensionality to stability[65]. If, say, we believed that if the dimensionality of space was other than three, then Newtons' laws would have been different or the gravitational potential would no longer be Poisson-like, the explanation under consideration would be non-starter.

---

bottom-up, the other perhaps non-causal and top-down, in terms of the equivalence principle. One has the sense that the two explanations are complimentary and do not compete. The relation between the explanation of *E* in terms of *X* and the explanation of *E* in terms of *T* does not seem like that. Instead, the putative explanation of *E* in terms of *T* seems redundant and superfluous, given the availability of an explanation in terms of *X*. This seems connected to our sense that there is something unsatisfactory about an explanation that postulates an unexplained correlation between *X* and *T*. I acknowledge that although this seems like a natural way to think about the example, it does not follow that this assessment is correct.

[65] Woodward, 2018 also expressed skepticism about whether there is any empirical way of ascertaining whether this claim about independence is correct. Thus the argument described above is a conditional one: if we can make sense of the appropriate independence claims, these provide a basis for the directional features of the explanation.

13. Conclusion

As noted earlier, many philosophers have attempted to connect asymmetries associated with causal direction with issues having to do with thermodynamic asymmetries, entropy increase the supposed need for a "past hypothesis" and the direction of time. The assumption seems to be that getting clear about these (broadly) "entropic" issues is required for an understanding of the directional features of causation. I certainly don't want to question the interest and value of developing accounts of these entropic issues. Nor do I claim that they have nothing to do with the independence features on which I have focused. On the contrary, I think the independence features are closely bound up entropic behavior. I want to suggest, however, that it is worth considering the possibility that the connection between causal and thermodynamic asymmetries may take a different form than is commonly supposed by philosophers. Rather than (or perhaps in addition to) thermodynamic/entropic asymmetries providing a sort of ground or basis for causal asymmetries (with the former being more fundamental) it may be instead that both asymmetries (thermodynamic and causal) at least in part derive from (or have a common source in) facts about independence and the absence of special kinds of tuning but where the most natural way of expressing these facts employs causal language[66]. To take one obvious connection, uncorrelatedness assumptions of various sorts have been used from Boltzmann (with his *Stosszahlansatz)* to contemporary authors (e.g., Myrvold, 2020) to explain facts about thermodynamic behavior[67]. Indeed, the same contributors to the machine learning literature discussed above have recently argued (Janzig et al., 2016) that the principle that the initial state of a physical system and the dynamical law governing it should be algorithmically independent (which is an algorithmic version of **VRI**) implies that the non-decrease of physical entropy for a closed system if entropy is identified with algorithmic complexity. In general, anti-entropic or anti-thermodynamic behavior is behavior that requires fine-tuning -- either of initial conditions in the sense of specific patterns of correlation among these or special tuning of these to dynamical laws. As I have attempted to explain, these also are the considerations that often underlie judgments of causal asymmetry.

References

---

[66] That is (recalling my discussion above) uncorrelatedness assumptions involve claims about *causally* independent factors, causes of a common effect tend (in the absence of some common cause of the former) tend to be uncorrelated, effects of a common cause are correlated and so on. In stating these assumptions we employ causal language.

[67] Also relevant here are the derivations of causal as opposed to non-causal behavior (effects occurring before causes) in linear response theory in Jackson (1995 and Nussenzweig (1972) discussed in Norton (2009) and Frisch (2014). These derivations make use of assumptions about the absence of highly tuned or anti-thermodynamic behavior showing how conclusions about reasonable causal behavior can be obtained from the assumptions and underlying laws. This is suggestive about the tight link between CSI and other anti-tuning assumptions and causal reasoning but there is no reduction of or replacement of the latter by the former.

Albert, D. (2000) *Time and Chance*. Cambridge: Harvard University Press.

Armstrong, D. (1997) *A World of States of Affairs*. Cambridge: Cambridge University Press.

Cartwright, N. (1983) *How the Laws of Physics Lie*. Oxford: Oxford University Press.

Craver, C. (2016) "The Explanatory Power of Network Models" *Philosophy of Science* 83:698-709

Daniusis, P., Janzing, D. Mooij, J. Zscheischler , J. Steudel , B. Zhang , K., Scholkopf, B. (2012) "Inferring deterministic causal relations" arxiv.org/abs/1203.3475

Davidson. D. (1967) "Causal Relations" *Journal of Philosophy* 64: 691-703.

Dowe, P. (2019) "The Direction of Causation" in Kleinberg, S. (ed.) *Time and Causality Across the Sciences*. Cambridge: Cambridge University Press.

Earman, J. (1974) "An Attempt to Add a Little Direction to 'The Problem of the Direction of Time'" *Philosophy of Science* 41:15-47.

Earman, J. (1976) "Causation: A Matter of Life and Death" *The Journal of Philosophy* 73: 5-25

Earman, J. (2011), "Sharpening the Electromagnetic Arrow(s) of Time" In. Callender, C. (ed) *The Oxford Handbook of Philosophy of Time*. Oxford: Oxford University Press, pp 485-527.

Eberhardt, F. and Scheines, R. (2007) "Interventions and Causal Inference" *Philosophy of Science* 74:981-995.

Farr, M. (2020) "Causation and Time Reversal" *British Journal for the Philosophy of Science* 71: 177–204.

Forester, M. , Raskutti, G., Stern, R. and Weinberger, N. (2018) "The Frugal Inference of Causal Relations" *British Journal for the Philosophy of Science* 69: 821- 848.

Frisch. M. (2014) *Causal Reasoning in Physics*. Cambridge: Cambridge University Press.

Hausman, D. (1998) *Causal Asymmetries*. Cambridge: Cambridge University Press.

Hausman, D. Stern, R. and Weinberger, N (2014) "Systems without a Graphical Representation" *Synthese* 191: 1925-1930.

Hempel, C. (1965) *Aspects of Scientific Explanation*. New York: The Free Press.

Hitchcock, & H. Price (Eds.), Making a difference. New York: Oxford University Press.

Hitchcock, C. (2018). "Causal models". In E.N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Hoover, K. (2001) *Causality in Macroeconomics*. Cambridge: Cambridge University Press.

Hoyer, P., Janzing, D. Mooij, J. Peters, J. Scholkopf, B. (2014) "Nonlinear causal discovery with additive noise models"

Jackson, J (1999) *Classical Electrodynamics*. New York:Wiley

Ismael, J. (2016) "How Do Causes Depend on Us? The Many Faces of Perspectivalism" *Synthese* 193.

Janzig, D. and Scholkopf, B. (2008) "Causal inference using the algorithmic Markov condition" arXiv:0804.3678.

Janzig, D., Mooji. J., Zhang, K., Lemeire, J., Zscheischler, J., Daniusis, P. Steudel, B. and Scholkopf, B. (2012) "Information-geometric approach to inferring causal directions" *Artificial Intelligence* 182-183: 1-31.

Janzig, D., Chaves, R. and Scholkopf, B. (2016) " Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference" *New Journal of Physics* 18 093052.

Lange. M. (Forthcoming) "Asymmetry as a challenge to counterfactual accounts of non-causal explanation" *Synthese*.

Myrvold, W. (2020) "Explaining Thermodynamics: What Remains to be Done?" Philsci-archive.

Norton, J. (2009) "Is there an Independent Principle of Causality in Physics?" *British Journal for the Philosophy of Science* 60: 475-86.

Nussenzweig, H. (1972) *Causality and Dispersion Relations*. New York: Academic Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufman.

Pearl, J. (2000) *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Peters, J., Janzig, D. and Scholkopf, B. (2017) *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge: MIT Press.

Price, H. (2007). "Causal Perspectivalism". In H. Price & R. Corry (Eds.), *Causation, physics, and the Constitution of Reality*. Oxford: Oxford University Press.

Price, H. (2014). "Causation, intervention and agency—Woodward on Menzies and Price". In H. Beebee, C. Hitchcock, H. Price (eds.) *Making a Difference*. Oxford: Oxford University Press.

Salmon, W. (1993) "Explanatory Asymmetry: A Letter to Professor Adolf Grunbaum from his Friend and Colleague" In Earman, J., Janis, A., Massey G. and Rescher, N. *Philosophical Problems of the External and Internal Worlds*. Pittsburgh: University of Pittsburgh Press.

Shimizu, S., Hoyer, P., Hyvarinen, A. and Kerminen, A. (2006) "Linear Non-Gaussian Acyclic Model for Causal Discovery" *Journal of Machine Learning Research* 7: 2003-2030.

Stern, R. (Forthcoming) "Causal Concepts and Temporal Order"

van Fraassen, B. (1980) *The Scientific Image*. Oxford: Oxford University Press.

Wallace, D. (2014) "The Nature of the Past Hypothesis" (talk at the Philosophy of Cosmology conference, Tenerife, September 2014).

Wigner, E. (1970) *Symmetries and Reflections: Scientific Essays*. Indiana University Press, Bloomington

Woodward, J. "Levels: What are they and what are they good for?"

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

Woodward. J. (2018a) "Some Varieties of Non-Causal Explanation" In Reutlinger, A. and Saatsi, J. (eds.) *Explanation Beyond Causation*. Oxford: Oxford University Press, pp 117-137.

Woodward, J. (2018b) "Laws: An Invariance- Based Account". In Patton, L. and Ott, W. (eds.) *Laws of Nature: Metaphysics and Philosophy of Science* Oxford University Press

Woodward, J. (2020) "Physical Modality, Laws and Counterfactuals" *Synthese*

Zhang, K., Zhang, J. and Scholkopf, B. (2015) "Distinguishing Cause From Effect Based on Exogeneity" arxiv.org/pdf/1504.05651