



Επιστήμη, Γνωσιολογία, Φιλοσοφία

THEORIA

ISSN 0495-4548 – eISSN 2171-679X

Responses

(Respuestas)

James WOODWARD*

Department of History and Philosophy of Science, University of Pittsburgh

* **Correspondence to:** James Woodward. Department of History and Philosophy of Science, University of Pittsburgh, 1101 Cathedral of Learning, 4200 Fifth Avenue, Pittsburgh, PA USA 15260 – jfw@pitt.edu

How to cite: Woodward, James (2022). «Responses»; *Theoria. An International Journal for Theory, History and Foundations of Science*, 37(1), 111-129. (<https://doi.org/10.1387/theoria.23513>).

Received: 2022-02-22; Final version: 2022-02-22.

ISSN 0495-4548 - eISSN 2171-679X / © 2022 UPV/EHU



This work is licensed under a
Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

I thank Katrina Elliott, Marc Lange, Jiji Zhang, Kun Zhang, Porter Williams, and Fernanda Samaniego for their comments. My responses follow.

Response to Elliott and Lange

Katrina Elliott and Marc Lange (hereafter EL) have many, many objections to my flagpole paper.¹ For reasons of space I will not address all of these but I will say that in my judgment their unaddressed objections are also unpersuasive.

One general issue raised by EL concerns the logic of the inference to conclusions about causal direction. EL depict me as holding that causal conclusions can be inferred just from information about relative frequencies (or at least they suggest that this may be my view²). They then ask: if relative frequencies just concern patterns of actual events, how can they be evidence for causal conclusions which presumably have modal import? There are several misunderstandings here—these have to do with the interpretation of causal modeling techniques generally and have nothing specifically to do with the techniques I describe in my paper. In particular, there are two different respects in which causal conclusions from causal modeling are *not* based just on information about relative frequencies. First, these techniques make use of information about *probabilities* (and information characterized in terms of probabilistic relationships like statistical independence). My understanding of probability is the standard measure theoretic one, described in Kolmogorov’s well-known axiomatization. (Of course the applications under discussion involve “physical” or “objective” probabilities,³ rather than probabilities understood as degrees of belief—I take this to be common ground between EL and myself.) For a number of well-known reasons, probabilities cannot be interpreted as or identified with claims about relative frequencies⁴ although information about relative frequencies can serve, in conjunction with other assumptions, as *evidence* for claims about probabilities.

This “in conjunction with other assumptions” clause is crucial—information about relative frequencies cannot by itself justify claims about probabilities. As a simple illustration, suppose that I draw balls randomly with replacement from an urn that is known to contain black and white balls and no other color. The result is information about the frequency with which white and black balls are drawn and statistical tests can be used to as-

¹ In what follows I refer to the paper which is the target of the responses (“Flagpoles anyone? Causal and explanatory asymmetries”) as “my paper” or “my flagpole” paper.

² The phrase “relative frequency” does not occur anywhere in my essay. Nor do I at any point advocate a relative frequency interpretation of probability. The principle that I call CSI is formulated as a claim about a relation between the absence of causal relationships and “statistical independence”. As noted below, by the latter I just mean “probabilistic independence”.

³ I won’t try to provide an account of “physical” probability but the reader can think about it in the following way: claims about physical probability can be assessed by standard statistical tests such as significance tests and frequency information can sometimes be evidence for physical probability claims. This contrasts with probability as degree of belief.

⁴ Among other considerations, relative frequencies are not countably additive and do not correspond to a sigma-field—basic elements of the Kolmogorov axiomatization. The strongest connections between relative frequencies and probabilities are given by the various laws of large numbers which definitely do not reduce probability claims to claims about relative frequencies.

sess the hypothesis (P) that, say, the probability that the next ball I will draw (on a random draw) from the urn will be white is 0.5. Claim (P) is not tested (or rejected or supported) *just* on the basis claims about relative frequencies. To support (P) it is also crucial whether the draws are random and with replacement—we make use of this information in our statistical test. In other words, it is the *combination* of relative frequency information plus information about the characteristics of sampling procedure—that it is random etc.—that gives us probabilities, rather than just facts about relative frequencies. Indeed, the claim that the draws are random is irreducibly probabilistic—it means that each ball has an equal probability of being drawn. So the form of the inference here is: probability information (e.g., that draws are random) + frequency information \rightarrow conclusion C about some other probability (that the probability of drawing a white ball is 0.5). Furthermore the \rightarrow in this inference should not be interpreted as a deductive entailment but instead should be understood in terms of whatever account of statistical testing one employs—e.g., error probabilities of accepting or rejecting C, within a classical (Neyman-Pearson) type framework.

A similar point holds when causal modeling techniques are employed. Although the underlying logic is not always made clear in this literature, the assumption (perhaps tacit) is that the relative frequency information that is used is the result of draws from some larger population via a sampling procedure that is random or at least such that it results in outcomes that are representative of the larger population. For example, when Mooij *et al.* (2016) test the reliability of the independent error models for inferring causal direction that I describe in my paper on data concerning the level of rainfall and altitude at various locations, they are assuming that their data are samples—random or at least representative—from some much larger population of rainfall/altitude pairs. This is what entitles them to think that they are working with information having to do with probabilities rather than just information about frequencies in a particular sample.

The preceding paragraphs describe one reason why thinking that inference to causal conclusions in causal modeling is a matter of inferring such conclusions from relative frequency information alone is a mistake. There is, however, another reason, which arguably is more interesting. The form of the inferences described in my paper is not: information about probabilities \rightarrow causal conclusions. The form is rather:

- (2) Information about probabilities **plus** causal information C1 \rightarrow other causal conclusions C2.

This is a point that I have stressed elsewhere—e.g. Woodward (2003)—and that is generally recognized in discussions of causation both in philosophy and in disciplines like statistics and machine learning. It is reflected in Cartwright’s (1979) slogan “no causes in, no causes out” which tells us that to infer to causal conclusions we need some kind of causal information as input (along with other information such as information about probabilities).⁵ We can think of claims of form (2) as “bridge” principles that establish

⁵ Another reason why it is obvious that we need such additional information of form C1 for causal inferences from information about probabilities is that information about probabilistic relations among variables radically underdetermines what causal relations obtain among these variables. Indeed this is so even if we assume generic connecting principles like the Causal Markov condition—these typically yield a large equivalence class of different causal models, for a given probability distribution, so that additional information is required for inference to a unique model, something that is by no means always possible.

connections between probabilistic information and causal claims. Consider the principle of the common cause (which is one simple consequence of the principle that I called CSI):

- (3) If (i) X and Y are probabilistically dependent, then (barring certain exceptions that are not relevant in the present context) either (ii) X causes Y or (iii) Y causes X or (iv) X and Y have a common cause.

This is an example of a connecting principle of form (2). Suppose I know that (3) holds and that (i) is true, and (ii) and (iii) are false. Then I can infer (iv). Thus I infer a causal conclusion (iv) from probabilistic information (i) and causal information—in this case the absence of causal connections of certain kinds, as indicated in (ii) and (iii).⁶ Obviously (3) and the fact that (ii) and (iii) are false are not just claims about probabilities, assuming (as I take to be common ground between EL and me) causal claims are not reducible to claims about probabilities.

The machine learning techniques for inferring causal direction described in my paper take a similar, although in some cases more sophisticated form. They use bridge principles that connect information about probabilistic relations (probabilistic dependence or independence etc.) conjoined with assumptions with causal content to make inferences about causal direction. The assumptions with causal content that are employed take a variety of different forms. For example, when it is assumed that when X and Y are statistically dependent and that (v) they have no common cause and that (vi) the causal relationship between them can be correctly represented by only one of two additive error models of form $Y = f(X) + U$, or $X = g(Y) + V$, with U and V error or noise terms, and (vii) that the model in which the noise is statistically independent of the candidate cause gives the correct causal direction, both (v) and the assumptions about possible functional forms made in (vi) are assumptions with causal content. These play, via (vii) an essential role in the inferences to causal direction that I describe. There are many other examples of such connecting principles that are employed in the causal modeling literature—these include the Causal Markov condition and various principles described by Jiji Zhang and Kun Zhang in their responses to my paper. As noted above, connecting principles can also make use of more specific causal information—e.g., that X does not cause Y .

Thus in response to the question posed by EL—how can information about what actually happens (e.g., in the form of relative frequencies) by itself support conclusions with modal content about what would happen (causal claims, counterfactual claims)—I agree that no such inference is warranted. However, as I have explained, the inferences to causal direction and to other causal conclusions described in my paper do not take this form. The “extra content” in the conclusions that goes beyond what actually happens derives from additional input beyond information about frequencies, with this additional input taking the form of both information about probabilities and causal information. In both cases this is modal information. Information about relative frequencies can be evidence for causal claims but only in conjunction with other non-frequency information.

⁶ Information about the absence of causal connections is certainly causal information. Indeed it is well-known in statistics and econometrics, that such absence information (“exclusion restrictions”) can play a powerful role in reliable inference to the existence of other causal relationships.

In a related discussion, EL write:

Statistical independence [as understood by Woodward], then, seems to be purely a matter of frequencies—of lack of “correlation”. So understood, though, CSI [my proposed principle relating causal and statistical independence] would seem to preclude fluky, purely coincidental, unlikely correlations among causally independent variables. (Elliott and Lange 2022, p. 58)

They express puzzlement about how I can claim otherwise as I do in my paper. But, as I have explained, I understand statistical independence (and dependence) as independence (dependence) *in probability* according to the usual definition: random variables X and Y are statically independent iff $\Pr(X, Y) = \Pr(X) \cdot \Pr(Y)$. Because statistical (in)dependence is defined in terms of probabilities it is not a notion that can be understood in terms of facts about relative frequencies. Understood in the way just described, nothing about statistical independence precludes “fluky coincidences”—indeed in non-trivial cases there will always be non-zero probabilities of these occurring. For example, if I toss a fair coin with the tosses being independent of one another, so that the tosses are i. i. d., there is a calculable non-zero probability of my getting a run of successive heads of arbitrary length which I take to be a clear case of a “fluke” or “coincidence”. That is just how probabilities behave.⁷

E and L also pose the following related objection:

[...] suppose that someone blew up various flagpoles over the course of a day in such a way that coincidentally, H and A for flagpoles when we observed them ended up being correlated. Then the directions of the causal relationships among H, S, and A would presumably be no different. But that directionality couldn't then be based on CSI. (Elliott and Lange 2022, p. 58)

First, I don't claim that inferences to causal direction can only be based on CSI. If, say, H , S and A are all statistically dependent, then you cannot use CSI based reasoning to infer causal direction, although you may be able to infer causal direction on some other basis. More fundamentally, I don't understand what EL mean by “coincidentally” in their example. A natural reading of their example is that H and A are statistically independent but that we are looking at coincidence like a long run of heads in independent coin tosses. In such a case CSI does apply. If on the other hand, the envisioned case is one in which H and A are probabilistically dependent, I don't understand what is meant by saying that the correlation EL describe is a “coincidence”—it is rather what we should expect given the dependence between H and A .

EL also consider the possibility that by statistical independence I have in mind an interpretation of probability in terms of chances. Of course if “chance” is just another word for physical probability in its usual measure-theoretic interpretation I'm happy to accept this suggestion. However, although EL do not explain what they have in mind by “chance”, I suspect that they intend something more elaborate. “Chance” as used by a number of contemporary metaphysicians is a notion that combines probability-like elements with ele-

⁷ I agree that if one thinks just in terms of relative frequencies it is not easy to see how to characterize a useful notion of probabilistic independence, but this is an objection to just working with relative frequencies and failing to distinguish these from probabilities.

ments that are quasi-causal or involve explanatory claims—the idea is that chances cause (or do something like cause) things to happen and/or perhaps that chances figure in explanations—e.g., of individual outcomes or frequencies of outcomes. This may be accompanied by the claim that existence of chances is something that we infer to via an “inference to the best explanation”. Of course this incorporates additional structure into the notion of chance beyond what is provided by the standard mathematical notion of probability. If this is what is intended, I would reject the suggestion that causal modeling techniques make use of chances, either as evidence for causal claims or in some other way.⁸ The additional structure above imported into the notion of chance is not needed to make sense of causal modeling techniques—indeed as nearly as I can see the additional structure does not fit well with the standard mathematical understanding of probabilities.⁹ My own view, for what it is worth, is that probabilities or probability ascriptions can be evidence for causal relationships and that some theories (like quantum mechanics) explain probabilities of outcomes but facts about probabilities (or chances) do not themselves cause or explain outcomes.

EL also raise a number of questions about the “metaphysical” status of principles like **CSI** and **VRI**. They ask: Are these “metaphysically necessary” principles connecting causation with statistical relationships, holding in all metaphysically possible worlds? They present me with a dilemma: either the (i) principles are metaphysically necessary or (ii) they are not. If I am claiming (i) I haven’t established this claim and it is dubious. (I fully agree, although this in part because I think the notion of metaphysical necessity EL employ is unclear—see below.) If I am claiming (ii) then the principles are at best contingent principles of (merely) epistemological significance for our world, although EL are perhaps skeptical even of that.¹⁰ In general they complain that I have not explained what causal direction is “based on” or “consists in”—something that they think of as requiring a metaphysical account of some kind. This complaint rests on the assumption (for which they don’t argue) that questions about what causal direction “consists in” are clear and well-posed and admit of illuminating answers (apparently in terms of “metaphysical necessities”) and that an account of causal direction should provide this. I see no reason to accept this assumption.

As I tried to explain in my paper, it was not intended as a contribution to a traditional metaphysical project aimed at explaining what causal direction “consists in” (at least in the sense of that phrase that EL seem to have in mind). At the same time I also suggested that

⁸ It is worth bearing in mind that the notion of causation assumed in standard causal modeling techniques is deterministic—the stochastic element comes in via the error term. These techniques do *not* assume a notion of causation that is chancy or probabilistic in senses of the notions of this sort that are assumed in the philosophical literature.

⁹ Thus in my view the assumption that there are only two possibilities—that physical probabilities must be interpreted either as relative frequencies or as chances (in what I called the elaborate sense) is mistaken.

¹⁰ At several points in their response, EL seem to suggest that if principles like **CSI** and **VRI** are not metaphysically necessary, they cannot reliably be used for causal inference even in the actual world—inferences to causal direction must be based on something else. Or at least in such a case we have no explanation of why **CSI** and **VRI** are useful in our world. This is an extraordinary claim. If these principles are contingently true in the actual world or for some set of systems in the actual world (or even usually but not always true) but not metaphysically necessary, why can’t we legitimately use them for inferences in circumstances in which they are true or usually true? Why isn’t that explanation enough for why they are useful?

the principles like CSI and VRI that I discussed were not of merely epistemological significance. EL are understandably puzzled about what I have in mind—I acknowledge that what I said was very sketchy. Don't metaphysics as they conceive of it and epistemology exhaust the possibilities? How could there be some other alternative?

Both because of space considerations and my own limitations I cannot describe in detail the alternative that I have in mind—I hope to do that more fully elsewhere.¹¹ Nonetheless let me try to say a bit more, while at the same time acknowledging that I have not yet figured out how to best express the ideas I am trying to describe. As will be apparent, my view is essentially the same as that described in Porter Williams' commentary.

My working assumption both in my flagpole paper and in other recent work such as Woodward (2021) is that human thinking about causation developed to be useful and to take advantage of various generic features F that are, as a matter of empirical fact, present in our world—these are what I called the “worldly infrastructure” that supports causal reasoning.¹² Our thinking about causation developed so as to exploit or make use of the presence of these features, both in the sense that the features are used in making inferences about or in learning about causal relationships (the epistemic side of things) but also in the sense that the way in which we understand causation and the applicability of causal thinking to the world is tied to the existence of these features. In (supposed) possible worlds—call them alien worlds—in which these features are systematically absent, the preconditions for the application of causal notions (that is, as we presently think about causation) are missing and hence what if anything we should think about causal relations in them is unclear. The anti-entropic world described by Williams is one example of a world in which the preconditions for the application of our causal notions are absent.¹³

¹¹ In particular in a co-authored paper with Naftali Weinberger and Porter Williams (Weinberger *et al.*, forthcoming) on which we are presently working.

¹² There are many other examples of worldly infrastructure in addition to those explicitly discussed in my paper. Other examples include the fact that interventions are often possible (rather than attempts at intervention always being confounded or otherwise unsuccessful), the fact that some variables are statistically independent of others, rather than everything being dependent on everything else, the fact that statistical dependencies in observational contexts are connected to true claims about the results of interventions and much else.

¹³ Here is a very rough analogy: a few decades ago there was extensive philosophical discussion of criteria for personal identity—in particular the relative roles of “bodily” and “psychological” continuity. To explore this philosophers proposed thought experiments in which minds were transplanted into new bodies, minds underwent fissions that resulted in the same mind being placed in two different bodies and so on, on the assumption that judgments about personal identity in such cases had determinate answers, which could then be used to elucidate our understanding of personal identity in the actual world. An alternative view, which is the one I favor, is that our thinking about personal identity is tied to various facts obtaining in our world—in particular that there are no such things as mind transplants, mind fission and so on. These facts are part of the worldly infrastructure that supports our thinking about personal identity. In supposed worlds in which these facts don't hold, there is no basis for determining which judgments about personal identity are correct. The empirical preconditions for the application of our thinking about personal identity are absent. Note that this claim need not be interpreted as implying that all claims about personal identity are false in such worlds—one might conclude this but one might also conclude instead that such claims lack determinate truth values. So also for judgments of causation.

In saying that preconditions for the application of causal notions are absent in alien worlds, we face a choice about what to say about claims about causal relations in such world. One possibility is to say that all claims about the existence of causal relations in alien worlds are false. Suppose that we adopt this possibility and consider **CSI** in the following formulation: (i) If X does not cause Y , Y does not cause X and X and Y do not have a common cause then (ii) X and Y are probabilistically independent. Then, under the first possibility, in an alien world, (i) will automatically be true and if (ii) is false, **CSI** will be violated. So in that sense **CSI** is “applicable” but false in alien worlds. A second possibility is to hold that in alien worlds, both claims about the existence and the non-existence of causal relationships lack truth values or at least that we have no basis for such claims. This presumably would lead us to say that we can’t apply **CSI** to such worlds. I’m not sure which possible description is best—the first may seem clearer but the second strikes me as somehow more natural. Perhaps either is appropriate as long as consistently followed. (As the reader will see I sometimes move back and forth between these two possibilities in what follows—a reflection of my struggle to clearly express what I am trying to say.) In either case, though, we can’t use **CSI** to reason about such worlds.¹⁴

The principles **CSI** and **VRI** described in my paper are just two examples of principles that exploit or make use of worldly infrastructure that supports causal reasoning. (There are a number of other examples of such infrastructure that I did not discuss). The presence of such infrastructure makes a world or a portion of it, “friendly” to causal reasoning and the applicability of causal notions; by contrast, when such structure is systematically absent we have a context which is unfriendly to causal reasoning in the sense that various empirical preconditions for the application of causal notions (or at least their useful or fruitful application) are absent.

What I had in mind by “minimal metaphysics” was the project of characterizing such infrastructure and explaining the role it plays in supporting causal thinking. It may have been a misjudgment on my part to describe this project as having to do with “metaphysics” in any sense (even minimal) and of course I have no desire to legislate about what counts as metaphysics. Part of my motivation for using this terminology was simply that the infrastructural features are “out there” in the world and metaphysics, broadly conceived, is supposed to be concerned with what is out there. Because these are worldly features an account of them and the role that they play in causal reasoning is not just an “epistemological” story about how we come to know facts about causal relations although the presence of the features helps to explain how such knowledge is possible. Let me add, though, that what matters is of course not the label we give to the infrastructure project but whether it is legitimate and important. I suggest that it is, whether or not we call it “metaphysics” and that it is distinct from the sort of metaphysics **EL** have in mind.

Although it is far beyond the scope of this response to discuss this issue in detail, it may help if I add that on my view an illuminating account of causal relationships and the infrastructure that supports reasoning about them should not be sharply separated from epistemological accounts of how we discover causal relationships—if one wants to talk in

¹⁴ In particular, for the sake of consistency, we shouldn’t claim both that there is no fact of the matter about whether causal claims are true or false in alien worlds and that **CSI** is violated (as opposed to being inapplicable) in such worlds.

terms of what causation “is” (very much not my preference) then the project of characterizing this should not be viewed as sharply independent of the project of characterizing the epistemology of causation—we should look for an account that relates these two enterprises.¹⁵ One reason for this is that we want an account of causation that is usable by us (“functional”) and this requires that we be able to determine in a substantial range of cases which causal relationships obtain. The possible ways we have of finding out about causal relationships thus help to shape our conception of what causation involves since (among other considerations) a usable notion of causation needs to be such that in a substantial range of cases we can reliably assess whether causal relationships obtain. A proposed metaphysics of causation according to which we can never reliably determine whether causal relations as characterized by that metaphysics obtain or according to which the standard procedures we have for determining which causal relationships typically fail to accomplish this because of the way causation is characterized metaphysically would, in my view, be highly problematic—I would be inclined to regard such a metaphysics as a non-starter.

Instead, our characterization of causation and the techniques we employ for inferring to causal conclusions should fit together in the sense that it should be intelligible why the techniques lead to reliable conclusions about causal relations so characterized. In the flag-pole paper I tried to accomplish this by, e.g. showing how the inferential techniques associated with additive noise models could be seen as answering questions about the outcomes of interventions, the latter of course characterizing causation within an interventionist framework. (The issue of why these inferential techniques “work” in the sense of delivering reliable conclusions about the results of interventions is also taken up in Jiji Zhang and Kun Zhang’s comments.) If one wants to think of the interventionist characterization of causation as “metaphysics” and the inferential techniques as having to do with “epistemology” these sorts of investigations seek to establish connections between the epistemology and metaphysics of causation. It seems to me that there is clearly a worthwhile project here, whatever one wants to call it. The viewpoint and projects just described contrast with EL’s apparent assumption that we should sharply separate epistemological issues from “metaphysical” ones and that what one says about causation should fit neatly into just one of these two categories, conceived as mutually exclusive and unrelated. The way in which they deploy their related strategy of framing the discussion around questions of whether the principles like **CSI** and **VRI** are metaphysically necessary or not incorporates this sort of separation and seems to me to leave little room for informative answers to “why do the inference procedures work?” questions of the sort I have just described. Indeed it often sounds as though the only answer EL can envision to such questions is an argument of some kind that **CSI** and **VRI** are metaphysically necessary.

With this as background, let me turn to some further details of EL’s discussion. EL claim that there “exists” a “possible world” in which two variables X and Y are statistically dependent but X does not cause Y , Y does not cause X and they have no common cause, in contravention of **CSI**. They infer from this that **CSI** is not metaphysically necessary—it is not part of the metaphysics of causation, properly speaking. EL do not provide any justification for their judgment that this scenario is possible. (Perhaps they

¹⁵ I say more in defense of this claim in Woodward (2021).

think they can tell by “intuition”.) As I said in my paper, I do not claim that CSI and similar principles are metaphysically necessary—if only because (as I said above) I doubt that “metaphysical necessity” as they understand it is a clear or useful notion. In general I think that their “I can imagine a possible world” argumentative strategy for reaching conclusions about causation (of any sort) is much more problematic (and question-begging) than they recognize. For starters if there is a metaphysically possible world in which CSI is violated on one occasion, it is hard to see why we should not also suppose that there are metaphysically possible worlds in which CSI is systematically violated—always or almost always. More generally, why not metaphysically possible worlds in which the Causal Markov Condition is always violated? And why stop there? How about worlds in which when one intervenes to change *X* and there is an associated change in *Y*, this is always because *Y* “just happens by coincidence” to change even though there is no causal connection between *X* and *Y*? If this is metaphysically possible, it presumably establishes that there is no “metaphysical” connection between causation and intervention, thus refuting interventionism as an account of the metaphysics of causation.¹⁶ Indeed it appears that this “thought experiment about possible worlds” methodology can be used to show that pretty much any attempt to connect causal claims with anything else (probability, intervention etc.) will fail to deliver metaphysical necessities concerning causation, since it can always be claimed that one imagine metaphysically possible worlds in which such connections fail.

One consequence is that insistence that an account of causation be framed in terms of metaphysical necessities (even putting aside other objections) seems to lead to a rather thin account of causation—one that omits much of what is of most interest about the concept, which has to do with how it connects with other concepts that we care about.¹⁷ But there is more. One additional obvious issue is how we reliably determine whether a judgment that such and such a world is metaphysically possible is “correct”. A related deeper issue is this: EL’s methodology seems to assume that built into our current ways of thinking about causation are commitments that tell us how to reliably apply that thinking in circumstances that are wildly different from those in which causal thinking developed. That is, it is assumed that our current ways of thinking about causation tell us how to apply causal notions in worlds in which CSI or various other candidate connecting principles fail, since it is on this basis that we make judgments of metaphysical necessity in *outré* worlds—judgments that, moreover, somehow tell us something important about the forms of causal thinking we employ to engage with our world. But why suppose that anything like this is true? On my view, our

¹⁶ Again, just for the record, I do not claim that interventionism is an account of the metaphysics of causation in EL’s sense. What I object to is their strategy of evaluating claims about causation in terms of what is metaphysically necessary, as established by thought experiments about possible worlds. Interventionism may be a mistaken account of causation but “I can imagine a possible world in which...” is not a good objection to it.

¹⁷ Of course it might be responded that there is nothing wrong with an account that describes such connections—these are just not part of the metaphysics of causation. But then why should we focus so much on the metaphysics of causation? Why use words like “crucial” to describe the role of this metaphysics in discussion of causation? Why privilege metaphysics of the sort EL have in mind as opposed to all sorts of other things that might be said about causation?

causal thinking is designed for (and developed in) a world in which certain generic empirical facts and connections between causation, probability and interventions obtain. The contrary view seems to involve an implausible kind of Platonism about our current thinking and concepts according to which these contains instructions for how to apply causal notions in circumstances that are bizarrely different from any that we ever encountered. This is a highly contentious view of how language and human thinking work.¹⁸ Instead I side with Porter Williams in holding that a defensible version of naturalism will hold that we simply don't know how to apply our current thinking about causation to such circumstances—again, the most defensible views are either that there are no causal relations in such worlds or that there is no fact of the matter about which causal judgments in such circumstances are correct.

Let us take stock. On the one hand, we can follow EL in focusing on whether various claims about causation including my candidate connecting principles are metaphysically necessary, where we assess these via judgments about what is true in metaphysically possible worlds. This leads to a sharp separation between the metaphysics of causation and how we find out about causal relations, since claims about the latter are, at least in most circumstances, not metaphysically necessary. Moreover, the infrastructure project as a distinctive enterprise drops out of the picture or at least is relegated to mere epistemology since it does not involve metaphysical necessities. One of many costs of this way of framing the issues is the apparent absence of any reliable way of determining whether the metaphysical judgments that, according to EL are crucial to providing an account of causation, are correct. By contrast, in pursuing the infrastructure project we avoid such arbitrary judgments. The infrastructure features themselves are ordinary if generic empirical features whose presence or absence can be ascertained by ordinary methods of scientific investigation. Relatedly, the strategies I describe for inferring causal direction which rely on those infrastructure features can be assessed for reliability by a combination of mathematical and empirical analysis—for example, one sometimes can *prove* that the strategies will have good reliability characteristics (error characteristics) if certain conditions are satisfied—good reliability characteristics in the sense of delivering reliable results about what will happen under interventions. In other cases one can assess reliability empirically by comparing the results delivered by the techniques with results about causal relationships that are known to be true on other grounds, as in the altitude/rainfall investigations referred to earlier. (We don't have to appeal to judgments about what is metaphysically possible in doing any of this.) The upshot is not metaphysical truths about causation but we do arrive at an elucidation of some of the worldly structures that are exploited in causal reasoning, and accompanying understanding of why causal reasoning works to the extent that it does, and along with this a treatment that locates causation within a wider web of other concepts. It seems to me that there is much to be said for this second project, especially since many aspects of it have not yet been well explored by philosophers. We should reject a framing of the issues around causation that does not allow us to see this second project as even a coherent possibility. Such a framing needs to be argued for, rather than just assumed or presupposed.

¹⁸ For a very different view of how language and thinking work—one that is much more in accord with my own view—see Wilson, 2006.

A couple of quick final remarks: EL suggest at one point that I may be thinking of **CSI** and **VRI**

as playing their epistemic roles because of the considerations that make one potential scientific explanation better (in the sense of “inference to the best explanation”) than another. (Elliott and Lange 2022, p. 56)

There is no reference to “inference to the best explanation” in my paper and this is not my view about the status of **CSI** and **VRI**—indeed, I am skeptical that inference to the best explanation, at least as understood by metaphysicians, is a valid inference form.¹⁹ If one wants a connection with explanation, **CSI** is perhaps more appropriately understood as disallowing inferences to “no explanation”: if X and Y are statistically dependent, then what is ruled out is saying that there is “no explanation” for this, where I interpret “no explanation” to mean that X does not cause Y , Y does not cause X and that X and Y do not have a common cause. Saying that we should accept the claim that there is some explanation for a statistical dependency over the claim that there is no explanation does not involve comparing competing explanations as to their “loveliness” or anything similar.²⁰

EL also say that I do not have grounds for rejecting claims about causal direction of the sort defended by Huw Price according to which causal direction somehow derives from facts about our perspective as agents unless I provide a metaphysical account of what causal direction consists in. I do not agree. If, as I claim, the grounds on which we infer causal direction are objective, non-agent-dependent facts like the presence of certain statistical patterns, I find it hard to see how subjective facts having to do with our perspective as agents can nonetheless be central to judgments of directionality or to what causal direction “is” (if one wants to talk that way).²¹ What’s the positive story about how that is supposed to work? EL seem to be assuming that how we find out about causal direction and what causal direction “consists in” can come apart in a really radical way.

Response to Jiji Zhang and Kun Zhang

I congratulate Jiji Zhang (JZ) and Kun Zhang (KZ) for their very lucid and informative responses to my paper. I will discuss both of their commentaries together since they make

¹⁹ I acknowledge, though, that the issues here are complicated and that a lot depends on how one understands IBE.

²⁰ I will add that **CSI**, **VRI** and the many other connecting principles employed in recent accounts of causal inference are far more specific and precise than generalities about inference to the best explanation and that unlike IBE (as noted above) there are various ways of assessing the reliability characteristics of many of these principles. I certainly don’t object to exploring possible relationships between the connecting principles and possible formulations of IBE, but I don’t think it is required to reinterpret the connecting principles in terms of IBE in order to show that they are legitimate.

²¹ EL may favor a sharp distinction between the bases on which we make judgments of causal direction and what causal “direction consists in” but this is not Price’s approach as I understand it. Price (very sensibly) wants to explain why we make the judgments that we do and he is no fan of the “what does it consist in” metaphysics that EL favor.

similar or related points. In my paper I was (among other matters) interested in the question of why the techniques I described for learning causal direction “work” in the sense of leading to reliable conclusions in some circumstances. I approached this question within a broadly interventionist framework, framing the issue as one having to do with how these techniques can tell us about the results of interventions (since this is what causal claims involve) even when the techniques are applied to “observational” data in which is not the result of deliberate experimental investigations. Focusing on the “independent error” techniques discussed in my Sections 6-9, a rough and simple version of my argument was that when a model can be found in which an observed third variable or an unobserved error term U is independent of a putative cause X but dependent on Y , then in the generic case, we can regard U as or as like an intervention on Y , and make use of the consequences of this: if X doesn’t change under such an intervention on Y , this is evidence that Y does not cause X . Further if, e.g., there are no common causes of X and Y , we may conclude that X causes Y . As JZ observes, I sometimes put this in terms of finding surrogates for interventions in observational data. Both JZ and KZ suggest that this understates the closeness of the connection between inference to causal direction in observational cases and cases in which there is deliberate experimental manipulation—they suggest that in observational cases one can typically find interventions (in the sense of exogenous sources of variation) in the data (and not mere surrogates for them) although information about these may not be obvious and statistical analysis may be required to uncover it. As JZ suggests, to the extent that this is true, this gives us a unified treatment of causal inference in both observational and experimental contexts—an outcome which I agree is very plausible and welcome. In general to the extent that it is correct that the techniques described by JZ and KZ are best understood as finding interventions or intervention-like signals in observational data, it is completely transparent why these techniques yield causal information, as this is understood within an interventionist framework.²²

JZ also suggests that the key element in the notion of an intervention, at least when this is used for purposes of causal inference, is the exogenous variation notion referred to above and that what matters is whether one can detect the presence of such variation in the available data, rather than whether all of the criteria for an intervention described in Woodward, 2003 are satisfied. The notion of exogenous variation is, as he notes, in some respects broader than my notion of an intervention. For example, he argues, following Steel (2005), that even if, in the scenario described above, the noise term U is not regarded as a cause of Y —hence not an intervention on Y in the sense of my (2003)—it is nonetheless true that if it represents a stochastic component in the generation of Y from X , this component can still be viewed as variation in Y that is exogenous with respect to X . JZ suggests that this can in turn be exploited in causal inference. In a related observation, JZ also notes that one can sometimes detect that there is some variation in Y that is exogenous with respect to X by means of patterns of dependence and conditional dependence in other variables, as is illustrated in his diagram 1b. In this case too, it is not required that one of the measured variables be an intervention on X in the sense of being a cause of X .

²² We thus have an illuminating answer to the question posed by EL concerning why the techniques work and why we are often entitled to rely on them.

Both for reasons of space and my own limitations I will not try to explore whether everything that can be accomplished by the notion of intervention in Woodward (2003) can also be accomplished by the notion of exogenous variation described by JZ but I'm happy to agree that the latter is an important notion in causal inference and that it is certainly in important respects intervention-like. I note though that in some cases (such as 1b), it appears that the identification of exogenous variation may require an appeal to additional principles such as faithfulness and it would be good to have a more systematic understanding of when this is the case.

JZ also raises the question of whether an apt characterization of the notion of intervention should satisfy a requirement of “mechanism preservation”. Consider a manipulation M that consists of pulling on the end of spring to extend it (measured by its length L), with M being exogenous and unconfounded. Within a certain range, the restoring force F exerted by the spring depends (we assume) linearly on L . But if we stretch the spring too much it will break—we will have destroyed the mechanism by which L influences F . Should we regard such a spring-breaking extension as an intervention on L or should we instead regard it as not an intervention (since it is not mechanism-preserving)? In my 2003, after struggling a bit with this question, I opted for the former alternative, in part because I was concerned that building a mechanism-preserving requirement into the notion of intervening on X with respect to Y threatened to introduce a kind of circularity if one wanted to connect the response of Y to interventions on X to whether X causes Y . I attempted to capture the case by describing it as one in which the generalization $F = -kX$ holds or is invariant under some range of interventions but not under others. The mechanism preserving alternative would presumably involve the idea that when extending the spring breaks it, this no longer counts as an intervention on L , so that one has a range of cases in which it is possible to intervene on L with respect to F (or at least possible to intervene in a way that preserves the $F = -kX$ relation) and a range of cases in which this is not possible because such manipulations would be spring-breaking. At the end of the day, though, I'm not sure how much one's choice about this (mechanism-preservation or not) matters.²³ What does matter, as JZ says, is that one recognizes that the mark of a causal relationship is that it is invariant under some (hopefully non-trivial) range of interventions and other changes. This in turn suggests an invariance-based strategy for testing for the presence of causal relationships—look to see whether there is evidence in the data that some hypothesized relation continues to hold under variations in the putative cause variable. This strategy and the assumptions about causal relationships with which it is associated (reflected in my principle **VRI**) have been fruitfully exploited in the recent literature on causal inference, as described both in my paper and in JZ's response.²⁴ My own view, for what it is worth, is that merely finding such a stable relationship in a body of data does not by itself conclusively show that the relationship is causal,

²³ Let me acknowledge a further problem, though, raised in comments by Jiji Zhang (private communication) and also briefly considered in my 2003. What about manipulations that, rather than breaking a mechanism connecting X to Y , create such a connecting mechanism where none previously existed? If Y changes under such a mechanism-creating manipulation of X , we don't want to conclude that prior to the manipulation X causes Y . There are possible strategies for dealing with such cases, with the details depending on how the cases are spelled out, but I lack the space for detailed discussion.

²⁴ See also Weichbald, S. and Peters, J. (2021) which describes a number of invariance-based causal inference strategies that may be used in cognitive neuroscience.

since for example, the variations in X over which $Y = f(X)$ is observed to be stable might all have been generated by an unobserved common cause of X and Y . (That is, there is not really exogenous variation in Y with respect to X .) Nonetheless the observation of stability is certainly suggestive evidence for a causal relationship, especially if the variation in value of X under which the relationship holds seems to be wide and to involve apparently different circumstances. I agree with JZ that when one concludes that the $X \rightarrow Y$ relationship is causal on the basis of such an inference one is in effect assuming that an intervention-like process on X with respect to Y is present in the actual data, which again gives us a kind of unification with the other inference procedures described above.

Turning now to KZ, he also emphasizes (as he has in previous work) the role that identifying the presence of interventions or intervention-like processes in observational data plays in causal inference. Again, the various results about inference to causal direction based on independent noise models described in Section 9 of my paper and to which KZ has contributed so centrally can be thought of as making use of this basic idea.

As noted both in my paper and KZ's response, the independent noise models discussed in his Section 4.1 rest on the assumption that there are no confounders between X and Y . Given this assumption, these models do not require an assumption like faithfulness. As KZ explains in his Section 4.3 it is possible to infer causal direction even when confounders may be present but this requires assumptions such as his AP1 and AP2 that are broadly similar to the more familiar faithfulness assumption employed in Spirtes *et al.* (2000). These assumptions are faithfulness-like in assuming the absence of certain kinds of "coincidental" deterministic relationships among hidden variables and in assuming the absence of structures that "hide" signatures of causal relationships in the observed statistics.

In a general sense these are "absence of special tuning" assumptions. They appear (*prima-facie* at least) to be different from the special tuning assumptions that must be satisfied in the anti-entropic world (but not the entropic world) discussed below in Williams' response (with the anti-entropic world satisfying the anti-tuning assumption I call **VRI**) but perhaps there are deeper connections. Obviously it would be very desirable to better understand the status of such assumptions and the conditions under which they hold. Anti-tuning assumptions are viewed with skepticism by many philosophers—they wonder why (or what guarantees that) the world should be such that these assumptions commonly or typically hold. For example, causal structures violating faithfulness in the sense of Spirtes *et al.* (2000) are certainly possible. My inclination is to think that such skepticism (when general) is the wrong perspective to adopt. Faithfulness and other anti-tuning assumptions are often plausible and one can sometimes although not always empirically detect violations. When they hold, they can legitimately be exploited in inferences. Assumptions concerning the absence of various sorts of special tuning are common in many areas of science and, as argued in Wallace (2019), it is not clear that science can proceed without reliance on some of them. Although I don't have space to argue this in detail, a great deal of causal reasoning in science including physics seems to rest on assumptions about the absence of special tuning of various kinds—a point that is suggested both by JZ's and KZ's discussions as well as that of Porter Williams, to which I now turn.²⁵

²⁵ Also suggested in Woodward (2016).

Response to Porter Williams

Porter Williams' (PW) extraordinary paper considers some of the implications of interventionism and assumptions like **VRI** for our thinking about causal relationships in an anti-entropic (or "time-reversed") world in which entropy globally decreases. As he notes such a world provides a striking illustration of the sort of possibility which **VRI** is designed to exclude. This is because to execute anti-entropic behavior particle trajectories have to be precisely tuned to the governing Hamiltonian, in contrast with an entropic world like our own in which this sort of tuning is not required. In addition **CSI** also appears to be violated (or alternatively, inapplicable) in a globally anti-entropic world because the initial micro-state will need to incorporate statistical dependence relations between, as he puts it, "the positions and/or momenta of particles with no antecedent causal connection".

After describing some of the highly peculiar features of such a world and the way in which a number of the generic infrastructure features (connected to **VRI**, **CSI** etc.) described in my paper fail in the anti-entropic world, Williams concludes that there is no basis for conclusions about which causal relations (that is, causal relations as assessed in terms of the way in which we currently think about causation) obtain in it. Too many of the conditions that underlie the applicability of causal reasoning in our world fail to obtain in the anti-entropic world and hence when confronted with such a world the most sensible reaction is that we don't know what to say about it, causally speaking. As he puts it, "epistemic humility demands that we simply withhold judgment about the presence or absence of causal relations in time-reversed worlds". This contrasts sharply with EL's confidence that they know which causal judgments are appropriate for worlds that are very different from our own. As I understand him, Williams thus opts for the second of the two possibilities that I presented in my discussion of EL—"don't know what to say" as opposed to "causal claims in the anti-entropic world are false". Of course I agree with the general conclusion that either there are no causal relations in a globally anti-entropic world or that we don't know what causal judgments are appropriate in such a world. A globally anti-entropic world is exactly the kind of world in which what I called the worldly infrastructure that underlies the applicability of causal notions fails to obtain.²⁶

As PW brings out, an anti-entropic world has many other features (besides failure of **CSI** and **VRI**) that are very peculiar or alien from the point of view of our world. Pretty much any local intervention in the time reversed world will transform it into a world in which entropy-increasing relations are present and these will spread, "infecting" more and more of the anti-entropic world. As a result a world in which local interventions are performed will not be a world which maintains its globally anti-entropic character. Or, to put it the other way around, the hypothesis that a world is and stably stays globally anti-

²⁶ Of course a globally anti-entropic world is still, by hypothesis, law-governed. But even if, every causal relationship is associated in some sense with a law of nature (cause \rightarrow law), the converse is not true within an interventionist framework—a law-governed world need not be one in which causal concepts apply or causal relations are present. Causal concepts have additional structure, as indicated by principles like **CSI**, the relation between causation and intervention, the directional character of causal relationships and so on. For example, the relationship between the entangled particles in an EPR-type experiment is law-governed but it is not causal, assuming an interventionist framework—one cannot intervene on one of the particles to affect the state of the other.

entropic implies that local interventions cannot be performed in that world. This in itself makes such a world seem causally uninterpretable in terms of an interventionist framework for understanding causation (or at least the framework is not applicable to such a world). Indeed if one of the marks of an agent is that it is able to manipulate its world with some degree of reliability, a globally anti-entropic world seems inconsistent with the existence of agents. But there is more. Putting aside the consideration just described, suppose that a macroscopic agent in a mainly entropic world attempts to implement an anti-entropic scenario—say, manipulating a fried egg so that it returns intact to the shell from which it came. Suppose, as seems reasonable, that there is a limitation on how fine-grained the interventions of such an agent can be (recall that the agent is macroscopic). Thus when the agent attempts to realize the very complicated set of initial conditions (characterized in a very fine-grained, micro vocabulary) I that will lead to an unscrambled egg, it is extremely likely that agent will instead realize one of the many initial conditions I^* that are close by to I and that will not lead to unscrambling. If M is one of the relatively coarse-grained manipulations the agent is able to perform then even if the micro-realizations of M include I , many, many more of those realizations will also include various versions of I^* . As a result, the relation between M and unscrambling will be highly unstable—even if there are some realizations of M that lead to unscrambling, the overwhelming majority of such realizations will not and the agent will not be in an epistemic position to detect in advance which realizations are which. Thus there will not (putting aside special cases in which the agent has an extremely fine-grained control over the micro-state of the system of interest) be an $M \rightarrow$ unscrambling relationship which is useful or exploitable for such an agent. Such agents will instead look for stable causal relations which, so to speak, follow the direction of entropy increase, rather than looking for relations that seek to undo entropy increase.

One of the many accomplishments of PW's discussion is that it brings out very clearly some of the connections between causation (and causal thinking) and the second law of thermodynamics. Such connections have been suggested or gestured at by many researchers but PW's commentary provides (at least for me) a new insight by elucidating how a world in which the second law is systematically violated is profoundly unfriendly to causal reasoning and the applicability of causal concepts as we ordinarily understand these. Our causal thinking is designed to fit a world which is very largely entropic.

Response to Fernanda Samaniego

I thank Fernanda Samaniego for her comments on my paper. I argued there that at least in many cases, causal directionality is not to be found just “in” the laws governing the behavior of a system but rather in the combination of (i) the laws, the initial and (ii) the boundary conditions characterizing the system (including what is fixed and what is allowed to vary) and the way in which (i) and (ii) interact with each other. I used examples involving the ideal gas law to illustrate this point. Samaniego argues that while this may be true for some cases it is not true in all—that is, in some cases it appears that the governing laws alone fix the causal direction in a system, in the sense that one cannot “reverse” that direction by considering a system with the same governing laws but different initial and boundary conditions. One of her examples is the flagpole case itself. The relationship between the height H of the pole, the angle A of the sun and the length S of the shadow, $S = H \cot A$,

seems to have a “baked in” causal direction running from H and A to S . Altering assumptions about initial conditions and which such conditions are fixed does not allow us to reverse the causal direction in the way that it does in the gas law example—or at least we cannot do this without changing the causal structure of the example in additional ways.²⁷ As best I can tell, this claim of Samaniego’s is correct. It thus becomes an interesting issue to try to characterize just what it is about the flagpole example that distinguishes it in this respect from some of the other examples I discuss. I will not try to speculate about this except to make the following brief remarks. First, the ideal gas law describes the behavior of a system at equilibrium and it is perhaps plausible that laws describing equilibrium relations will not have a built-in causal direction, with the causal direction in any particular system to which the law applies instead depending on facts about the initial and boundary conditions governing the system. Second, fundamental laws (putting aside complications having to do with the weak force which are not relevant to the sorts of examples under discussion) are time-reversal invariant. This seems to suggest that when we model systems in terms of such laws any causal directionality present in these systems is not going to come from the laws alone but will also need to have its source in facts about the initial and boundary conditions governing the system. The generalization that we use to explain the length of the shadow in terms of the flagpole height is neither a generalization describing equilibrium behavior nor a time-reversal invariant fundamental law so the sources of directionality associated with initial and boundary conditions that are operative in these cases are not present in the flagpole example. This perhaps lends strength to the suggestion that the directionality in the example comes just from the law. However, even in this case, information about the exogenous determination of H somehow seems relevant to our judgments about direction. We know that the value of H is caused (indeed fully fixed) by whatever processes are involved in the manufacture of the pole and I take this to be inconsistent with the claim that S causes H . Samaniego suggests that the difference under discussion derives from the fact that H produces S but that such productive relations are not present in the ideal gas example. I would encourage her to say more to elucidate the notion of production involved in these judgments.

REFERENCES

- Cartwright, N. (1979). Causal laws and effective strategies. *Nous* 13, 419-437.
- Elliott, K., & Lange, M. (2022). Running it up the flagpole to see if anyone salutes: A response to Woodward on causal and explanatory asymmetries. *THEORIA. An International Journal for Theory, History and Foundations of Science*, 37(1), 53-62 (<https://doi.org/10.1387/theoria.22351>).
- Mooij, J., Peters, J., Janzig, D., Zscheischler, J. and Scholkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research* 17, 1-102.
- Steel, D. (2005). Indeterminism and the causal Markov condition. *The British Journal for the Philosophy of Science*, 56, 3-26.

²⁷ Of course one might imagine a system in which a sensor measures the length of the shadow and then adjusts the angle of the pole over the course of the day so that the length of the shadow remains constant. Here S causally influences A . But this is a fundamentally different system than the one that is usually considered.

- Spirtes, P., Glymour, C. and Scheines, R. (2000). *Causation, prediction and search*. Cambridge, MA: MIT Press.
- Wallace, D. (2019). Naturalness and emergence. *Monist* 102, 499-524.
- Weichbald, S. and Peters, J. (2021). Causality in cognitive neuroscience: concepts, challenges, and distributional robustness. *Journal of Cognitive Neuroscience* 33, 226-247.
- Weinberger, N., Williams, P. and Woodward, J. (Forthcoming). The Worldly infrastructure of Causation.
- Wilson, M. (2006). *Wandering significance: an essay on conceptual behavior*. Oxford: Oxford University Press.
- Woodward, J. (2003). *Making things happen: a theory of causal explanation*. New York: Oxford University Press.
- Woodward, J. (2016). Causation in the sciences. In Paul Humphreys (ed.), *The Oxford Handbook of Philosophy of Science*. New York: Oxford University Press, 163-184.
- Woodward, J. (2021). *Causation with a human face: normative theory and descriptive psychology*. New York: Oxford University Press.

JAMES WOODWARD is a Distinguished Professor in the Department of History and Philosophy of Science at the University of Pittsburgh.

ADDRESS: Department of History and Philosophy of Science, University of Pittsburgh, 1101 Cathedral of Learning, 4200 Fifth Avenue, Pittsburgh, PA USA 15260
Email: jfw@pitt.edu