Loose Talk Kills: What's Worrying about Unity of Method

Nancy Cartwright*†

There is danger in stressing commonalities among methods because the differences matter in fixing the meaning of our claims. Different methods can, and often do, test the same claim. But it takes a strong network of theory and empirical results to ensure that. Failing that, we are likely to fall into inference by pun. We use one set of methods to establish a claim and then draw inferences licensed by a similar-sounding claim that calls for different methods of testing. Our inferences fail, and bridges we build (or policies we set) depending on them fall down.

1. What's Here. Perhaps we can find a sufficiently general and abstract description that homogenizes our many disparate scientific methods. But there is danger in the enterprise. Failing a good, well-established, thick theory of the concepts we employ, the methods we use to test claims involving these concepts play a central role in fixing exactly what those claims are. And we must be clear about exactly what claims we are making if we are to know what inferences we can draw from them. We must be wary of 'inference by pun': drawing inferences that follow legitimately from a claim employing the same words but that is not the claim we have established, whose meaning is fixed in part by the test procedures we have used to establish it. I elaborate this worry in section 2. Section 3 illustrates using causality as an example. But the lessons apply equally to a great many central scientific concepts.

2. What Claim Are We Making? Two decades of emphasis on scientific practice in science studies underline that a vast number of practices make up

*To contact the author, please write to: University of Durham, Department of Philosophy, 50 Old Elvet, Durham DH1 3HN, UK; e-mail: nancy.cartwright@durham.ac.uk.

†Thanks to Alex Marcellesi for help with both ideas and production.

Philosophy of Science, 83 (December 2016) pp. 768–778. 0031-8248/2016/8305-0011$10.00
Copyright 2016 by the Philosophy of Science Association. All rights reserved.

https://doi.org/10.1086/687862 Published online by Cambridge University Press

science, including classifying; experimenting; measuring; settling on standards; designing and building machines (e.g., machine physics) and technology (like the laser or Kelvin's Atlantic cable); constructing models and blueprints for real-life situations; creating new substances and materials and new ways to change the world (which contemporary genetics witnesses repeatedly); developing new mathematics; discovering, creating, and stabilizing phenomena; calculating, inferring, making, refining, and defending concrete claims about specific situations and specific systems; providing a vast number of low-level ceteris paribus laws; and so on. It is this thick web of procedures and applications that give content to our scientific claims. Without them the claims are mere words, as Thomas Kuhn argued:

> Consider . . . the quite large and diverse community constituted by all physical scientists. Each member of that group today is taught the laws of, say, quantum mechanics, and most of them employ these laws at some point in their research or teaching. But they do not all learn the same applications of these laws, and they are not therefore all affected in the same ways by changes in quantum mechanical practice. On the road to professional specialization, a few physical scientists encounter only the basic principles of quantum mechanics. Others study in detail the paradigm applications of these principles to chemistry, still others to the physics of the solid state, and so on. What quantum mechanics means to each of them depends on what courses he has had, what texts he has read, and which journals he studies. . . . Though quantum mechanics (or Newtonian dynamics, or electromagnetic theory) is a paradigm for many scientific groups, it is not the same paradigm for all. (Kuhn 1962, 49–50)

These ideas are developed by Peter Galison, who argues that different scientific groups imbed what can seem the same concepts in very different networks of practice and inference that thus embody different understandings. Groups with different 'thick' understandings communicate to some extent via stripped-down 'pidgin' languages—but the understandings and practices imbedded in these pidgins are far from sufficient to do the tasks of any of the separate groups (Galison 1997).

Kuhn's points generated decades of discussion about the rationality of theory change, supposing that he had shown that new theory and old do not refer to the same thing with the same words and hence do not contradict each other. My concern is rather with the use to which we put scientific knowledge. The topic in this symposium is unity of method. Maybe there are some things all scientific methods have in common, and there may be some good purposes served by adumbrating these. But when it comes to use, it is crucial to stress the disunity and particularity of the methods and practices that interpret and support scientific claims. There has long been the

https://doi.org/10.1086/687862 Published online by Cambridge University Press

rosy hope that we can establish results that sail free of the methods, practices, techniques, successful predictions and applications, and rich web of concrete low theory that interpret and support them: we can take scientific results and put them on a shelf in a knowledge supermarket for consumers to take away to use in new homes. This is dangerous. If the uses we make of scientific claims are to be successful, we must ensure that the methods and practices used to establish a scientific claim are sufficient to support the inferences we draw from it. To do the contrary is to fall into science by pun.

The claims of science must be supported—in detail—by empirical facts. This support is witnessed by success in predicting and intervening precisely in the world. If so, what is supported are claims as interpreted through the network of concrete assumptions and practices that afford the successful predictions and interventions. Although I shall illustrate my worries about unity of method by drawing out the lessons of causal pluralism, there is nothing special about causality; what is true of it is widely true. It is important to keep different methods differentiated because they so often are methods appropriate to finding out about different things, even if these things share a common name; what follows from knowledge about one of these things will not follow from knowledge of another.

Figure 1 illustrates differences between monist and pluralist accounts of scientific concepts, using 'cause' for illustration. Causation is what Otto Neurath called a *Ballung* concept, a concept with rough, porous boundaries, a congestion of different ideas and implications that in various combinations are brought into focus for different purposes and in different contexts (see Cartwright et al. 1996; Cartwright and Bradburn 2011). Concepts like this can, and often do, play a central role in science, and especially in social science. But they cannot do so in their original form. To function properly in a scientific context, they need to be made precise.

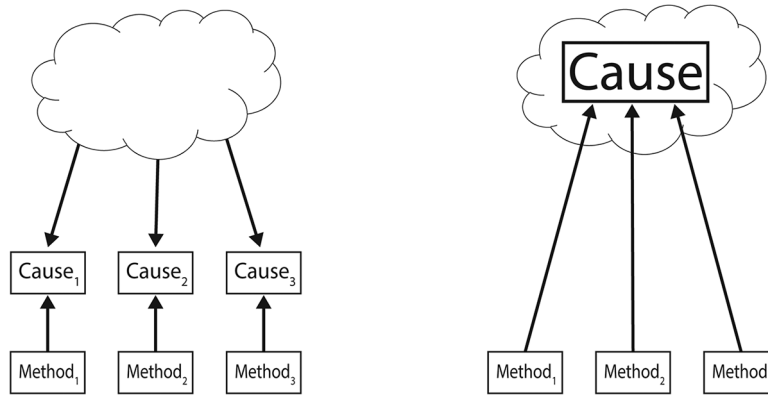


Figure 1. Causal pluralism vs. causal monism.

https://doi.org/10.1086/687862 Published online by Cambridge University Press

The right-hand side of figure 1 pictures a central, nonfuzzy core to this concept that can, possibly with great effort, be characterized precisely; 'causes' picks out some one relation, and the variety of methods employed in discovering causes are all methods for testing for that one relation. We establish this relation in a variety of ways, but no matter how it is established, we can draw the same conclusions.

Despite vast philosophical effort, we have not found any such central core that can properly be used in science, where rigor matters. The notions on offer and our theories about them are too thin; they do not allow us to draw many conclusions beyond what follows narrowly from the assumption that a given relation satisfies our definition. We don't have an empirically well-supported web of inferences that attribution of the defined relation supports.

The left-hand side of figure 1 fits better with what I see in scientific practice. 'Cause' has no precise 'central core'. The trick in bringing it into a scientific context is not to find the 'right' characterization, but rather to construct a characterization that does the jobs that we want to do when we use the term in that context. When it comes to testing causal claims, different methods are appropriate to different characterizations—and correlatively, for different characterizations, different inferences can be drawn from claims employing the concept 'cause'. Characterizations will be done in different ways in different scientific disciplines, serving different ends and to fit with the different concepts, methods, assumptions, and standards operating in these disciplines. The more precise scientific concepts that result are then different from each other and different from the original *Ballung* concept.

I sometimes use the ugly word 'precisification' to describe the process by which a *Ballung* concept is transformed into one fit for science. Sophia Efstathiou (2009) calls this process "found science" on the analogy of found art. Damien Hirst's shark in formaldehyde is still a shark, but it is not the same shark as when it was swimming in the sea. It has been made suitable for an artistic context, to serve specific artistic purposes. In Efstathiou's words, the shark has been "founded"—given a form appropriate to serve its new purposes—in the artistic context. But the shark "founded" as art loses most of its original functionings, including its ability to be founded in other contexts, such as shark soup.

Similarly, the different foundings of a *Ballung* concept like 'cause' or 'race' are not different ways of characterizing what remains the same concept. What can be shown true of 'cause' under one founding cannot be presumed true under another, and empirical methods that work for telling where one obtains do not normally secure a causal relation that has been founded in any other way. In particular, they do not license inferences that follow given other foundings.

The left-hand side of figure 1 pictures causal pluralism: different concepts differently founded for different purposes, all equally entitled to the

https://doi.org/10.1086/687862 Published online by Cambridge University Press

name. Causal pluralism is not a philosopher's game; it has real bite. We must not establish causal claims that have one interpretation and then use them as if they had a different one, or we are headed for disaster. As both Popper and the positivists insisted, where scientific claims are to be taken seriously, it matters exactly what those claims are.

Consider evidence-based policy, where three different kinds of causal claims are regularly conflated: (*a*) generic claims that hold locally: a cause-kind causes an effect-kind in some particular kind of situation ('C causes E somewhere'); (*b*) generic claims that C causes E 'widely'—across a range of situation types; (*c*) singular predictions that C will cause E in a specific new situation ('C will cause E here').[1] Here is an example, not untypical but particularly striking because it mixes all three together, without note, in just one sentence, from a defense of randomized controlled trials (RCTs) in development economics by Esther Duflo and Michael Kremer, the first an MIT economist, the second at Harvard, both at the Jamil Poverty Action Lab: "The benefits of knowing which programs work . . . extend far beyond any program or agency, and credible impact evaluations . . . can offer reliable guidance to international organizations, governments, donors, and . . . NGO's beyond national borders" (Duflo and Kremer 2005, 205).

'Which programs work' is about what works to produce the targeted effect widely, impact evaluations tell us that a program produced the effect somewhere, and 'reliable guidance' predicts that the program will produce the effect in a new situation. This kind of conflation can lead to bad policy predictions, and bad policy predictions can lead to bad development outcomes, completely contrary to the hopes of the authors.

The worry is that policy will proceed by pun. A program is implemented somewhere and a careful after-the-fact evaluation is carried out to see whether it achieved the intended results, say, via an RCT. This can show that the program had a positive effect size—for example, a positive difference in the mean of the targeted effect in treatment and control groups—in the situation where it was tested. Add a handful of metaphysical and empirical assumptions, including that the situation of the treatment and control groups is governed by a set of the same linear causal laws, that there is a probability measure over the quantities in those causal laws for the combined treatment/control population, and that the study design ensures that the treatment and control groups differ only by values for the treatment variable and its downstream effects (or nearly enough so). Then a positive effect size shows that the treatment genuinely figures as a cause of the effect in a law governing that situation there (see Cartwright and Hardie 2012).

This provides evidence that the same cause figures in a law for a new situation only supposing that the new situation is governed by the same rel-

1. Of course, these kinds of claims are themselves multiply ambiguous.

https://doi.org/10.1086/687862 Published online by Cambridge University Press

evant causal law as the study situation. And it provides evidence that the effect size will be positive in a new situation only if,[2] in addition, the interactive causal factors in that law have the same distribution in the target as in the study situation—an assumption that is often unlikely to be true. So a great deal of work needs to be done if the methods that establish 'It works there' are to be relevant to 'It works here', and you cannot avoid this work by just eliding the 'here' and 'there' and acting as if the claim established and the claim put to use are one and the same.

Of course, the problem is not restricted to concepts of causality but is endemic throughout the sciences. Nor am I alone in my concerns. William Wimsatt, for example, issues a similar warning against punning in science: "Applying a heuristic to a problem *transforms* the problem into a non-equivalent but intuitively related problem. . . . But it is not the same problem, so beware: answers to the transformed problem may not be answers to the original problem" (2007, 135).

3. Some Foundings for 'Causes'. There are a variety of different accounts of causality prevalent in philosophy of science now. Each lays out important facts about how a system of relations must behave if the relations are to be labeled 'causal' under that particular precisification; each can be seen as a distinct way of founding more precisely the concept of causation. I shall focus on two: (1) a notion of causality suited to James Woodward's level invariance test, and (2) a notion characterized by the three basic axioms for causal Bayes nets (CBNs): (i) the causal Markov condition (CMC), (ii) faithfulness, and (iii) minimality.[3]

(1) Woodward (2003) introduces *level invariance* in the context of what I call "epistemically convenient linear deterministic systems" (ECLDSs; Cartwright 2007, 161). A linear deterministic system (LDS) is *epistemically convenient* if each effect in the system has at least one special cause that does not cause anything else in the system except by causing that effect, and the set of special causes is variation-free.[4]

Suppose we take—as in Cartwright (2007, 154), for instance—a set of familiar-looking axioms like irreflexivity and asymmetry that a scientific concept of causation might satisfy, plus the assumption that the causal laws

2. Almost 'only if'. Effect size is a function of the mean, which can of course be the same without the same distribution.

3. Roughly, CMC says that causal parents screen off effects from everything else except effects of the effect; faithfulness, that causal parents and offspring are probabilistically dependent; and minimality, that there are no more causal relations than necessary to account for the probabilistic dependencies and independencies described under CMC and faithfulness (see Spirtes, Glymour, and Scheines 2000).

4. That is, together they can take any combination of values from their separate ranges.

https://doi.org/10.1086/687862 Published online by Cambridge University Press

of an ECLDS are responsible for any other true functional relations that obtain in it (i.e., that all 'spurious' relations derive from genuine causal laws) as an implicit definition—in Efstathiou's words, a 'founding'—of a concept of causality that could obtain in an ECLDS. It is then provable that testing for Woodward's level invariance is a reliable method for arriving at causes in the sense of 'cause' picked out by these axioms: an equation satisfies Woodward's level invariance criterion if and only if it is a causally correct principle in this particular precisification of 'cause'.[5]

(2) Let me now turn to CBNs. Here we find theorems that show the tight fit I advocate among characterization, methods, and use. Many of the CBN algorithms for discovering causal claims are provably valid so long as the three central CBN axioms are satisfied.[6] That means we can take the axioms to characterize what is meant by 'cause' and be assured that the methods for finding out about causes are reliable for finding out about relations thus characterized as causal.

We also are assured that we have reliable ways to put that causal knowledge to use. There is, for instance, a well-known "manipulation theorem" in Spirtes et al. (2000, 47) that describes what happens under a 'manipulation' of a given quantity to the probabilities of other quantities, assuming that these are all governed by causal principles that satisfy the CBN axioms relative to the joint probability over the total set of quantities. But 'manipulation' is a very specific kind of change under very specific assumptions about the operative causal principles, including their stability as values of variables change. So warranting the claim that some particular set of actions we perform count as manipulations will be no easy matter, and we must be careful not to be misled by casual use of the term 'manipulation'.[7]

This tight fit of method and use to characterization has a downside, however. The methods are appropriate only for the concept of causality specified in the characterization, and similarly for the inferences we draw from causal claims, for instance, about what will happen under manipulation. Wolfgang Spohn, I take it, would be happy with this. He is explicit in maintaining that causality is characterized by CBN axioms: "Bayesian nets are all there is to causal dependence," as the title of his essay claims (Spohn 2001). But that does not seem to be how all advocates of the methods see it.

5. The proof also provides a much-needed characterization of what it means for an equation with a dummy variable representing 'missing factors'—like the regression equations Woodward and others discuss that play a central role in social and economic sciences—to be causally correct.

6. Some are also valid for specific weakenings and others for specific strengthenings of the axioms. What I say in the text follows for those as well, assuming that the characterization of 'cause' is adjusted accordingly.

7. Glymour himself is careful about this. See Meek and Glymour (1994, 1010).

https://doi.org/10.1086/687862 Published online by Cambridge University Press

Consider, for instance, a repeated conversation between Clark Glymour and me, which can help to make clear the points I have been urging and will bring the discussion back to the start. I worry about cases where the specified axioms fail. Glymour responds by showing all the wonderful ways that have been developed (and some that people have long been using) to cope in those cases to draw causal conclusions without relying on the axioms. But what then are we making claims about? How are we to understand what we mean when we claim 'X causes Y'? Glymour seems uninterested in characterizations of 'causes', and so far as I know no characterization has been offered as such since the first work by Glymour, Scheines, and Spirtes (which seemed to adopt my view that generic causal claims are claims about what singular causal happenings would occur with what probability—inviting, of course, a characterization, in turn, for the concept of singular causation).[8]

Perhaps, though, I am illicitly supposing causal pluralism. Perhaps all these different methods, the ones where the axioms hold and those that cope when some of the axioms fail, are—as on the right-hand side of figure 1—just different methods of finding out about a single causal relation, and in particular one for which the inferences we wish to make hold. In that case my worry about drawing inferences by pun is unfounded. But still the overall question looms. What guarantees that these methods are correct for establishing claims that support our inferences? No matter whether a term bears a monist or a pluralist interpretation, in science, where rigor matters, we need arguments to justify that the methods used for establishing claims support the conclusions drawn from them. There may be a kind of unity of method here: our methods for testing claims with the word 'cause' in them may all be targeted at one and the same relation. But without a good characterization of what that relation is and how it behaves, we are not entitled to suppose that.

This is a worry I have about overoptimistic application of the nice work of Judea Pearl (2000). I worry about the comprehensiveness of his methods, not their validity. Pearl offers a complete methodology from hunting causes to using them. First, he provides a general way to represent causal principles, maintaining that his representations are general enough to treat any kinds of causal principles we are familiar with. I don't quarrel with this here. Second, he offers a detailed semantics for inferring singular counterfactuals from causal principles of this form. Nor do I quarrel with this here. Third, he points to reliable methods like CBNs and RCTs for inferring causal

8. This is also the view supposed in the Holland and Rubin defense of RCTs, which they take to provide the average individual effect size (see Holland 1986).

https://doi.org/10.1086/687862 Published online by Cambridge University Press

principles from probabilities. Though we probably disagree about how widely the assumptions hold that are necessary for these methods to be valid, I agree that the methods are powerful and can be reliable. The scheme is ideal. We have trustworthy methods for going from data to model and from model to prediction. So the predictions are well supported by the empirical evidence.

The problem is in the join-up. We need reasons to suppose that the causal principles that produced the data in the studied situation are the same as those that produce the outcomes we want to predict in the target situation. But we seldom have such guarantees. The probabilistic methods that Pearl and others endorse for discovering causes can provide good descriptive accounts of the network of causal relations that obtain in various populations. These can be a part of the evidence base for the more basic science that allows us to predict what the causal principles might be in new situations.

But simple induction, even if the models are for what is supposed to be the 'same' population, is seldom a good tool of inference—and to be warranted in using it, we need good reasons to believe we are studying an entrenched structure. Otherwise, for new situations we need to predict new principles, and we can't do this by collecting statistics on populations in the not-yet-existing situations. To predict laws for new situations, we need theory and the large and tangled confluence of evidence and hypotheses that go into building up and supporting reliable theory. This is well beyond the network of statistics and background empirical and metaphysical assumptions that matter immediately to correct use of CBN methods or applications.

4. Conclusion. Just what can we do when we know that X causes Y? As I have stressed, that depends on what 'causes' means in this claim, and that, in turn, is tightly tied to how the claim is established. So rather than focus on unity of method, I urge that we devote more attention to comparisons among our methods and the various precisified concepts they measure. Sticking with the *Ballung* notion of cause for illustration, we need answers to the following:

1. Which foundings of 'cause' are various familiar methods for causal inference good for discovering/testing?
2. How good are they at these jobs, both in the ideal and in the kinds of conditions that obtain in practice?
3. How good is the argument that vindicates each method? Are there deductive proofs as there are for some CBN algorithms? Do the arguments depend on empirical assumptions? If so, do we have a good, well-grounded sense of the kinds of situations in which those assumptions obtain?

https://doi.org/10.1086/687862 Published online by Cambridge University Press

4. What kinds of inferences follow given causal claims understood under various foundings? What kinds of strategies for achieving what we want are justified?
5. How good are the arguments that show that these inferences are valid, with the same questions about empirical assumptions as in item 3?
6. What features of concrete situations or specific problem settings can help us recognize which foundings, which methods of causal discovery, and which uses of causal knowledge are likely to be suited to them?

This last item is one of the big jobs facing us: getting concrete. Our characterizations are generally too abstract. For example, I have praised work in CBNs for its formal definitions and proofs. These guarantee that we are not guessing about what methods are valid for discovering CBN causality, or for making further inferences from CBN-interpreted causal claims, or for when they are applicable. The explicit assumptions and definitions tell us when. But what do these conditions amount to in the real world? That is issue 6 in my desired catalog of comparisons, and it is of the utmost practical importance.

I return at last to the general lesson my discussion of 'causes' is meant to elaborate. The topic for this symposium is unity of method. Perhaps, as some of the panelists urge, the methods it is appropriate to use for measuring and testing in the sciences can be unified in some way; perhaps we can find some core, some way in which they are all the same. That's fine so long as we do not lose focus on what makes them different, for it is the precise details of the differences that make them fit for the job of scientific testing—that ensure they mesh precisely with the kind of precisely characterized concepts that we employ in science to produce precise and reliable conclusions.

Method must march hand in hand with the characterizations that our axioms and definitions lay out. There can be no more unity to the one than to the other. Philosophers still hold out hopes for a single sense of central concepts like 'causes', or perhaps for limiting the splintering to two or three senses—maybe, for 'causes', some version of a production account and some version of a difference-making account. I laud these attempts. It would be very helpful to have a single account, or at least a single core that all acceptable accounts have in common. Still, we want our sciences to produce exact, unambiguous, and precise claims. For these purposes a loose account of what we are talking about or a highly abstract characterization or a mere loose common core will not suffice. We will inevitably be using different foundings for different purposes. What matters is to ensure that the founding fits the purpose and that the methods we use for causal inference are good enough to warrant the conclusions we draw. Inference by pun is un-

https://doi.org/10.1086/687862 Published online by Cambridge University Press

warranted inference. It is loose talk, and, as the World War II injunction warns, loose talk can kill.

REFERENCES

Cartwright, Nancy. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

Cartwright, Nancy, and Norm Bradburn. 2011. "A Theory of Measurement." In *The Importance of Common Metrics for Advancing Social Science Theory and Research: Proceedings of the National Research Council Committee on Common Metrics*, 53–70. Washington, DC: National Research Council.

Cartwright, Nancy, Jordi Cat, Lola Fleck, and Thomas E. Uebel. 1996. *Otto Neurath: Philosophy between Science and Politics*. Cambridge: Cambridge University Press.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. Oxford: Oxford University Press.

Duflo, Esther, and Michael Kremer. 2005. "Use of Randomization in the Evaluation of Development Effectiveness." In *Evaluating Development Effectiveness*, ed. George Pitman, Osvaldo Feinstein, and Gregory Ingram, 205–32. New Brunswick, NJ: Transaction.

Efstathiou, Sophia. 2009. "The Use of 'Race' as a Variable in Biomedical Research." PhD diss., University of California, San Diego.

Galison, Peter. 1997. *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press.

Holland, Paul. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Meek, Christopher, and Clark Glymour. 1994. "Conditioning and Intervening." *British Journal for the Philosophy of Science* 45 (4): 1001–21.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Spohn, Wolfgang. 2001. "Bayesian Nets Are All There Is to Causal Dependence." In *Stochastic Causality*, ed. David Costantini, Maria Carla Galavotti, and Patrick Suppes. Stanford, CA: CSLI.

Wimsatt, William C. 2007. *Piecewise Approximations to Reality: Engineering a Philosophy of Science for Limited Beings*. Cambridge, MA: Harvard University Press.

Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.

https://doi.org/10.1086/687862 Published online by Cambridge University Press