Book review: reduction of surprise: some thoughts on Dowding's conception of explanation

Reduction of Surprise: Some thoughts on Dowding's Conception of Explanation[1]

Keith Dowding's book "The Philosophy and Methods of Political Science" is an exciting tour de force through some important debates in the philosophy of the social sciences. One aspect I admire, in particular, is the philosophical pluralism it displays. The book encourages us to reflect on the nature of our discipline by showing us ways to think about our methods, without being too philosophically prescriptive. It provides guidance for political scientists reasoning about the foundations of their work. The book reflects the breadth of debates in the philosophy of the social sciences and provides orientation in the thicket of "isms", theories and paradigms. Of course this wouldn't be Dowding's book if it wasn't opinionated.[2] But it is also a book that shows the philosophical and methodological choices that are available to the political scientist and theorist, and it endorses more than one view.

In this short comment I focus on the role of explanation and how (if at all) causation is or should be related to explanation in political science. This leads me into a brief detour around the philosophy of models before I return to causation and the notions of mechanisms and process tracing.

***

Chapter 3 tackles one of the most intractable issues in the philosophy of science: what is an explanation, and by virtue of what is an explanation a good explanation? Dowding answers both questions by talking about the importance of recognizing patterns and making the world more predictable. In a slogan: "explanation is the reduction of surprise". What does that mean? I think we are supposed to read this in a functional sense: the function of explanations is to reduce surprise.

Dowding says that the reduction of surprise means that the same explanation applies to relevantly similar settings:

> "… any claim that 'X' explains 'Y' means that under the relevant circumstances X will always explain Y." (p. 43)

One important role of an explanation is to provide reasons for an outcome. So we are less surprised by the world if we have a good explanation because the same explanation gives us reasons that explain many different phenomena in relevantly similar circumstances. I call this the **explanatory power desideratum**. One could also say that the best explanations do a lot of explaining for us; they are explanatorily fecund.

But alternatively one can also read "reduction of surprise" as a surprise reduction in terms of predicting outcomes. That would mean that a good explanation is predictive (or retrodictive) in relevantly similar situations and thus reduces our surprise about what will happen in many situations. I call this the **predictive power desideratum.** An explanation has predictive power if it gives us many reliable predictions.[3]

---

1 I would like to thank Peter John for organizing the roundtable and symposium and two anonymous referees for their helpful comments.

2 For readers who do not have the pleasure to know Dowding and his work I should explain that Keith was once perceived to be a staunch defender of rational choice approaches and formal methods. This perception was probably always wrong, as Keith's methodological commitments and interests have consistently been much more far-reaching. This book reflects this (yet again) clearly.

3 I have been prompted by a referee to point out that the level of surprise depends on the expectations we have, and that we cannot be surprised if we do not have expectations. Dowding could respond that if we do not have any

I suspect that Dowding wants both desiderata: a good explanation (i) is supposed to be explanatorily fecund, and (ii) offers predictive (or perhaps retrodictive) power about an expected course of events in relevantly similar situations. Note that the distinction between the two desiderata can become important. It helps us realize that explanations are more than just reliable predictions, a point also endorsed by Dowding. To see this, consider a "magic predictor" who miraculously predicts the course of future events. But also suppose that the internal workings of the magic predictor are a black (magic!) box to us. The magic predictor provides us with great predictions (by assumption) but is a total failure as a provider of explanations because she fails to give us any reasons why things are happening.

It is less clear what to say about the inverse case. It appears that for many explanations prediction is a litmus test – without some successful predictions the explanation is poor. But arguably that is not true of all explanations. Explanations of unique historical events provide reasons but arguably do not predict. (Though there is a lot of disagreement about this claim: perhaps even these seemingly singular explanations are quasi-predictive for relevant counterfactual scenarios.) Perhaps computational models of complex systems are also like that – many of them are not good at making predictions, but they give us a good sense why certain outcomes can come about. They have high explanatory but low predictive power.

A similar issue arises when we consider freak events, i.e. events with very low probabilities. Suppose there is a skydiver surviving a double parachute failure (a bit of googling reveals somewhat similar real-world cases). It is possible to come up with ex post explanations for this freak event. But I am not convinced that one would be able to identify a pattern that reduces surprise, or that one would be able to make predictions. If that is true, then it seems that at least some explanations are not surprise reducing in the predictive sense. Sometimes there are plausible ex post accounts of a singular event that are explanatory.

The upshot of the discussion so far is that the relation between the two aforementioned desiderata is usually tight, but not necessarily so. Not all explanations also serve as good predictions, and not all predictions explain well. In particular, prediction is not sufficient and perhaps not even necessary for explanation.

The prediction requirement is also problematic for mathematical explanations, especially for proofs, though for quite different reasons. A proof shows that a statement is true, typically by offering a series of derivations (often identity statements) until a statement is reached that is obviously true. Mathematical proofs are, in effect, tautologies: they are true by logical necessity. There is little doubt that proofs are explanatory. Showing the proof step-by-step is often illuminating and helpful. But are proofs predictions in Dowding's sense? Once there is a theorem we do not predict that certain mathematical relations hold, we know that they *hold necessarily*, not as a prediction but as a logical truth. Dowding is very explicit about his desire to count mathematical models as explanatory devices, but I struggle to see how they meet his prediction requirement for good explanations.

Perhaps my nit-picking here is exaggerated. It could be that a mathematical proof is predictive in a particularly strong sense: it is a prediction that is necessarily true, in the same sense that Pythagoras' theorem is true in Euclidean space. However, personally I think that that way of thinking about predictions is a stretch. In the next section I will make a suggestion how Dowding could have avoided this problem.

---

expectations, everything surprises us (in a certain sense of surprise). I believe Dowding takes surprise as the portion of unexplained variance.

One thought about related literature. Dowding's take on explanations could potentially be enriched by looking at recent attempts to think of explanation in terms of unification (Kitcher 1981). The unification thesis says that one good-making feature of explanations is that they explain many different phenomena in a parsimonious framework. That seems quite close to Dowding's view – a connection that could be explored in greater detail.

***

One area that Dowding's book could cover in more detail is the philosophy of science analysis of *model use*. Models are peculiar beasts. We all use them all the time, but they are hard to pin down, philosophically. Dowding takes models to be "sets of statements related formally or analytically to generate testable hypotheses" (p. 80). Also, "[a] good model is isomorphic to that which it represents in the relevant aspects" (p. 79-80). They are simplified representations of the target system, and the simplification helps us to understand the relevant aspects of the target system better.

The relation between model and target system is subject to an ongoing debate within the philosophy of science. However, it seems to me that models in the social sciences are rarely isomorphic to their target systems. That would require a one-to-one mapping between all elements of the target system and the model, preventing the required simplifications, abstractions, and idealizations. Granted: Dowding only claims an isomorphic relation between "relevant aspects" of the target system and the model. But to determine the "relevant aspects" we already need a model of the target system, so that proposal sounds circular.

The wider point here is that the relationship between models and target system is probably not one of isomorphism, but rather one of similarity. A model is an adequate model if it represents some aspects of the target system. But how it does that can vary depending on the type of model employed. As Sugden (2000) points out, many models are "credible worlds", fictitious stories that share some resemblance with the target system. Sugden uses Akerlof's "Market for Lemons" (1970) as an example. Of course the real market does not look anywhere close to Akerlof's toy market for "lemons." Still, the real market and the toy market share structural features that make them similar enough to find the toy market useful for explaining what is happening in the real market.

This leads to another important point. Models are objects, not propositions. Primo and Clarke (2012; following similar debates in the philosophy of science, see Giere 1988; Teller 2001; Godfrey-Smith 2006) use maps as an example. A map is not a set of propositions or statements. It is a drawing on a piece of paper. The crux is: since maps are objects (and not statements) they cannot be true or false. Maps are a particularly vivid example because they are physical objects. But the point holds more generally. Take a computer model. It consists of programming code. The code of a computer model is neither true nor false (it can be buggy, but not false). The model can be more or less useful for explaining the behaviour of the target system because it is more or less similar in a relevant sense to the target system. It is the similarity relation that makes the model good or bad, not that it is true or false.

Dowding disagrees: "… we easily predicate truth-values to maps precisely because, like language, they are signals carrying meaning" (p. 81). I am not convinced: a map, to my mind, is neither a signal nor a proposition. One might derive propositions from maps (for example: directions like "From Aldgate East you need to walk 200m to the West to reach Shoreditch High Street", which can be true or false), but the proposition is created by the interpreter of the map, it is not already entailed by the map. Similarly, a dynamic computer model or a rational choice model provide us with the trajectory of a process through time or the behaviour of an idealized rational agent. By

itself, the model is neither true nor false. Once we begin to interpret the output of the model as a prediction about what the target system does, then these predictions might be true or false, but not the model itself. The model is just more or less useful.

The debate about models in the philosophy of science is far from over, and perhaps Dowding's more traditional account turns out to be the right one and my preferred account turns out to be wrong. However, my view comes with a benefit: If Dowding were to subscribe to it, he would have a much easier time accounting for the explanatory role of mathematical proofs and theorems. I argued above, somewhat pedantically, that theorems do not predict. But perhaps theorems, at least when appealed to in the social sciences, have a dual function. They are theorems – but they are also models! Take the Condorcet Jury Theorem as an example. In the first instance, the Condorcet Jury Theorem is what it says on the box: a theorem. But when it is discussed in the social sciences it takes on a second role: it does not only feature as a theorem but also as a model of a democratic voting process. It features as a model because there is a similarity relation between the theorem's votes and real votes, and the theorem's decision making procedure and real-world decision making procedures. It is not the set of true statements that makes the theorem interesting as a model, it is the similarity relation.

The advantage of the dual role view of theorems is that we can now account for our intuition that theorems are explanatory. Strictly speaking, the mathematical theorem as such does not explain anything – it just states a complicated mathematical tautology. But the theorem as a model does explain: it explains, for example, why in voting processes in which voters are competent and their votes independent the majority is likely to vote for the correct alternative. In addition, this explanation has the two desiderata of explanatory and predictive power. The same explanation holds and the same outcome is predicted in relevantly similar cases.

***

In this final section I link the notion of explanation with the notion of causation. Is there something special about scientific explanation? Quite often it draws on making causal claims. It is tempting to think that all scientific explanation is ultimately causal, but we have already seen examples above where that is not so obvious. Nevertheless, many paradigmatic examples of successful explanation are based on accounts of causal relations.

Here are two different types of causal explanation social scientists are typically interested in:

- An explanation why a particular event happens;
- An explanation of repeatable patterns.

For instance, we can ask why the Copenhagen Climate Conference failed. What do we mean by that question? Most often we want to know why a conference of the <u>type</u> of Copenhagen fails. That is what interests us from a predictive point of view. Good answers help us to avoid repetition. However, sometimes we want to know why the <u>token</u> Copenhagen conference fails. We will then not be happy with explanations that refer to a type-based explanandum ("I did not ask why such conferences typically fail – I want to know what went wrong with this one!").

The difference between these two different requests for explanation might be rooted in two very different understandings of causation, as explained by Ned Hall (2004, p. 225):

> "Causation, understood as a relation between events, comes in at least two basic and fundamentally different varieties. One of these, which I call "dependence", is simply that:

> **counterfactual dependence** between wholly distinct events. In this sense, event c is a cause of (distinct) event e just in case e depends on c; that is, just in case, had c not occurred, e would not have occurred. The second variety is rather more difficult to characterize, but we evoke it when we say of an event c that it helps to generate or bring about or produce another event e, and for that reason I call it "**production**."

In vintage Dowding-style, the second form of causation features as "bump-bump" causation in the book. This is a nice suggestive term for a rather elusive idea – that causation must in some sense be transferred in a physical process from one object to the other (see Dowe 2008 for a summary). Dowding takes a pluralist line with regard to theories of causation. Causal explanations are taken as narratives, and the quality of narratives depends on the needs and questions of the audience. I prefer to take a more pronounced stance on this. I want to suggest that, at least as far as the social sciences are concerned, the adequate account of causation is an account of "counterfactual dependence", an account of causation-as-difference-making. The core point of this notion of causation can be expressed in a sentence: If C were to occur, E were to occur; and if C were not to occur, E were not to occur.

Causation-as-difference-making is the notion of causation we appeal to when we conduct experiments, run regressions, use matching techniques, and so on. It is often said, however, that difference-making accounts are less suitable when we want to give an account of the kind of causation we're after when doing case studies, process-tracing, looking at mechanisms, etc. Like Dowding, I want to resist the idea that the bump-bump analysis of causation really is the adequate account of causation to underpin process-tracing, mechanisms, and so on. Also, again concurring with Dowding, I doubt that there really is such a fundamental difference between large-n and small-n methodology. (At least when it comes to philosophical foundations – there might be a lot of difference in practice.)

To go back to the Copenhagen conference: If the demand is for an explanation of the outcome *of this token conference*, then we shift down the granularity of the analysis. We identify robust causal relations not between structural features and outcomes, but between more-fine-grained phenomena. For instance, we often go on the level of intentional agency. Since we know a lot about how intentional agents typically behave, we have implicit large-n knowledge of robust causal relations at that level. We then end up with chains of these more fine-grained causal relations. This might be a form of process-tracing or mechanism analysis, but it is still based on the same notion of causation: causation-as-difference-making. There is no need to appeal to a mysterious bump-bump account. I believe that Dowding and I are in agreement here since Dowding says that the reductions required are not from type to token, but from type to a different type-level of causation with finer granularity.

Dowding characterizes the more fine-grained (mechanistic) accounts of a larger phenomena as "narratives". What I dislike about the metaphor of narratives is the implicit suggestion that the quality of the explanation/narrative is in the eye of the beholder. Is this really that subjective? Dowding (p. 145) claims that any given counterfactual is based upon large-n generalization. But often this happens (at most) in implicit style (in the form of general psychological experience, etc.). My own take is that we are only beginning to understand the quality standards for more micro-based explanations in the social sciences. I doubt that these standards will be primarily psychological or dependent on the needs and desires of the audience. My guess is that the robustness of mechanisms will be important, and that we will develop more transparent standards for assessing the quality of such explanations as our ability to employ such explanations improves. But this is a question for another time.

***

Keith Dowding has written an inspiring book. Its ambition should not be underestimated: while the last few decades have seen some pretty good philosophy of economics, the philosophy of political science is in its infancy. If it is beginning to grow up now then this is to some extent due to Dowding's excellent parenting. I am immensely grateful to Dowding for putting the philosophy of political science on the agenda with his great book and I hope that this is the beginning of many fruitful debates.

References

Akerlof, George A. 1970. "Market For Lemons -- Quality Uncertainty And Market Mechanism." Quarterly Journal Of Economics 84(3): 488–500.

Clarke, Kevin, and David M Primo. 2012. A Model Discipline: Political Science and the Logic of Representations. New York: Oxford University Press.

Dowe, Phil. 2008. "Causal Processes." In: The Stanford Encyclopedia of Philosophy (Fall 2008 Edition), ed. Edward N. Zalta. Available at: https://plato.stanford.edu/archives/fall2008/entries/causation-process/.

Giere, Ronald N. 1988. Explaining Science: A Cognitive Approach. Chicago: University of Chicago Press.

Godfrey-Smith, Peter. 2006. "The Strategy of Model-Based Science." Biology and Philosophy 21: 725–40.

Hall, Ned. 2004. "Two Concepts of Causation." In: Causation and Counterfactuals, eds. John Collins, Ned Hall, and Laurie Paul. Cambridge, MA: MIT Press, 225–76.

Kitcher, Philip. 1981. "Explanatory Unification." Philosophy of Science 48(4): 507–31.

Sugden, Robert. 2000. "Credible Worlds: The Status of Theoretical Models in Economics." Journal of Economic Methodology 7(1): 1–31.

Teller, P. 2001. "Twilight of the Perfect Model Model." Erkenntnis 55(23): 393–415.