LSE CPNSS

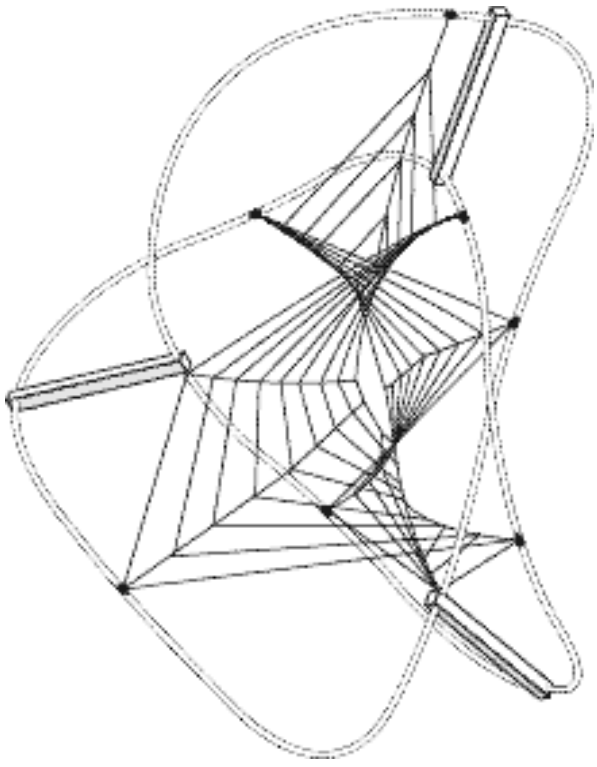
Centre for Philosophy of Natural and Social Science

Causality: Metaphysics and Methods

Technical Report 09/03

From Causation to Explanation and Back

Nancy Cartwright



Editor: Julian Reiss

From Causation to Explanation and Back*

Nancy Cartwright
Department of Philosophy, Logic and Scientific Method
London School of Economics
and
Department of Philosophy
University of California-San Diego

December 2002

*Corresponding address: CPNSS, LSE, Houghton St, London WC2A 2AE, philcent@lse.ac.uk. Research for this paper was supported by a grant from the Latsis Foundation as well as a grant for a research assistant from UCSD, for both of which I am extremely grateful. I would also like to thank Jordi Cat, Roman Frigg, Julian Reiss and Christoph Schmidt-Petri for their help.

1 Introduction

The dominant topics in general philosophy of science during the second half of the 20th century were surely *the nature of scientific theory, realism* and *explanation*. The discussion of *the nature of theory* today is still very much in the same terms as that of 50 years ago although the central paradigm has changed. The standard view, to which I will return below, used to be that a theory is a deductive system formulated in a formal language.[1] Nowadays it is probably the 'semantic view' that dominates: theory is a set of models and any language will do for describing exactly what the models are like.[2] Nevertheless the basic presumptions are the same. We ask the same question, and in asking it we presume both that it is a worthwhile question and that it has an answer – that there is some form shared across the sciences that a good theory should take. The methods of adjudication remain the same too: internal coherence and goodness of fit with what we deem to be our best scientific theories, with the principal cases now, as then, drawn from physics.

The nature of *realism* debates has shifted more. Under the influence of the Logical Positivists, for a long time Anglophone philosophy of science eschewed metaphysics. We did not talk about the truth of scientific theories nor about the reality of scientific entities but rather about scientific progress or the accumulation of scientific knowledge. The presumption was that whatever it is we are looking for when we evaluate a scientific theory, the characterisation must be internal to our own activities; it should not depend on reference to relations between theory and the world. So we discussed how effective theories are at problem solving[3], or whether they produce from within themselves, relying on their own heuristics, corrections that yield new predictions and applications.[4]

Nowadays we have forsaken our Positivist prohibitions and talk wantonly of truth, near truth, resemblance to the truth, aiming for the truth, getting the structure right, getting the theoretical content right, true predictions and, within the semantic view of theories, of resemblances or isomorphisms between models and the world. Recent versions of the semantic view may seem more guarded.[5] They talk not of isomorphisms or resemblances between models and reality but instead, between models and other models, usually data models or models of the phenomena. Nevertheless there is the tacit assumption that these latter are themselves supposed to be accurate, true, true enough or what have you. The language and the issues of the debates of the 1960's and 70's have been entirely replaced.

Still, it is in the third area, that of *explanation*, that discussion has shifted most, for here the topic itself has changed. We no longer talk about explanation; its place has been taken by *causation*. We do of course still discuss explanation

[1] The most extensive and sophisticated treatment can be found in Carnap 1956.

[2] Early references include Suppes 1967. Classical References include: Suppe 1989 and Giere 1988.

[3] Laudan 1996, Part 3

[4] Lakatos 1970

[5] van Fraassen 1980

itself, but it is not at the core of the field. Probably it appears most centrally as a tool in the realism debates. We are entitled, many suppose, to regard as true, the *best explanation* of a given body of phenomena; this behooves us to provide an account of what a good explanation is. This situation is a real change. For three decades every undergraduate course in philosophy of science started out from the Deductive-Nomological account of explanation. Nowadays, large numbers of philosophers of science do not engage with the topic at all.

This essay will be concerned with this dramatic shift. I shall trace the gradual transformation of *explanation* into *causation* and outline where I think we are now and where we could profitably be heading.

2 From causation to explanation

Carl Gustav Hempel's *Deductive-nomological* account of explanation[6] dominated work in philosophy of science for thirty years, and it still influences a number of related issues today. We should begin with a puzzle. Why was explanation thought to be such an important matter? Explaining is after all only one of a great many activities that the sciences engage in. Theories are used to explain, but they are also very importantly used to effect changes in the world. Yet theory application and use has never been a central topic. We also expect our theories to be well-confirmed if we are to rely on them, but during the last half century confirmation has certainly taken a backseat to explanation. Scientists also approximate, they classify, they build models, they borrow methods and ideas and techniques from elsewhere, they measure and they experiment. Each of these activities needs a philosophical treatment – when is it done well, when badly and how can we tell; and each has its own philosophical problems. Why did explanation stand out?

I want to focus on one particular function that explanation served that gave it a special status. For decades explanation stood in lieu of causation – causation, which is critical if we are to account for how science can change the world. This role is particularly notable because the situation is now reversed; for better or for worse we now tackle questions about both the methods and the metaphysics of causation directly.

David Hume taught that talk of causation was metaphysics and should be consigned to the flames. The Logical Positivists similarly scorned metaphysics, and for the most part for them too this included causation. But Rudolf Carnap had a way to salvage some important metaphysical concepts.[7] He urged that there are many concepts that appear to be in the 'material mode' but that would in fact be more perspicuously expressed in the 'formal mode'; that is, there are many concepts that seem to be directly about the world but that are really instead about our own theories and descriptions of it.

Various concepts of causation came to be treated in just this way. Kepler's laws hold on account of Newton's. This fact is a central part of the argument for

[6] See Hempel and Oppenheim 1948 and Hempel 1966.
[7] Carnap 1928

believing in Newton's laws. But how can one law be responsible for the obtaining of a second? We also regularly cite one event as the cause of another. Hempel's well-known example is the drop in temperature that caused the radiator to crack.[8] But, following Hume, what can this amount to over and above spatio-temporal connection and regular association between the two event types?

Explanation is the answer. Newton's laws explain Kepler's and the drop in temperature explains the cracking of the radiator. In both cases it is the same notion of *explanation* that does the job, and it is a notion that refers entirely to relations within language. The fact to be explained is deducible from a set of well-confirmed law claims and initial conditions, where the specific factors cited in explanation are factors that we find particularly salient among these. That is the famous *Deductive-Nomological* (D-N) account of explanation I mentioned above. It is important to bear in mind here that for the Positivists and their followers, explanation was not a mark in language of a relation that holds in the world but rather that the explanatory relation in language *is* the relation that holds; we only mistakenly project it onto the world.

3 From explanation to causation

The Deductive-Nomological model was recognised from the start as too restrictive. Insistence on deduction from strict laws would rule out a large number of explanations that we have high confidence in. Hempel added the *Inductive-Statistical* (I-S) model:[9] explanations may cite law claims and individual features from which it follows that the facts to be explained are highly probable even if they do not follow deductively. The underlying idea in both cases is that the factors offered in explanation (the *explanans*) should provide good reason to expect that facts to be explained should occur (the *explanandum*). This harkens back to Hume's account[10] in which the idea of causation is a copy, not of an impression of a relation in the external world, but rather a copy of an impression of our own feeling of expectation that the effect will occur when we observe the cause. For Hume this feeling of expectation results from our repeated experience of the cause being followed by the effect. For advocates of the D-N and I-S models, the expectation must be rational, not habitual – it must be justified under well-confirmed law claims.

The natural response in this case is to amend the requirements for I-S explanation. The *explanans* need not render the *explanandum* probable; it need only raise the probability. But that will not do either, as Wesley Salmon's example of probability-lowering explanations makes clear.[11] Salmon considered a case in which exactly one of two radioactive elements is randomly inserted into a box. Imagine that the strongly-radioactive element has a probability of .9 of disintegrating; the weak element disintegrates with probability .1. The probability

[8] Hempel 1942
[9] Hempel 1962
[10] Hume 1997/1748, Sections 4-7
[11] Salmon 1971, pp. 62-65

of decay when the weak element is in the box is thus far lower than that when it is not (since when the weak element is not in the box, the strong one is). Yet if decay occurs when the weak element is in the box, it is undoubtedly the presence of the weak element that explains it, and this despite the fact that the weak element lowers rather than raises the probability of decay.

Salmon's proposal was that the *explanans* should be *statistically relevant* to the *explanandum* – the probability of the *explanandum* should be different when the *explanans* obtains from when it does not. But that does not work either, as we can see by amending Salmon's set-up slightly: use two structurally different radioactive elements that are equally strong in disintegrating. Whichever element is in the box when a decay product occurs will explain that occurrence even though the probability of decay is the same with both elements.

Why does the cause – a genuine positive cause – counter-intuitively lower the probability of the effect in Salmon's set-up? Because the presence of the weak cause is correlated with the absence of an even stronger cause. This suggests following the strategy common in the social sciences: stratify the background population. That is, divide the population of trials into subgroups one contains all the trials on which the strong element is presented; the other, those in which it is absent. In a population of trials in which in every case the strong element is present, introducing the weak element will increase the probability of decay, and the same is true in a population in which the stronger element is universally absent. Following this line of thought, the natural proposal is to insist that the *explanans* increase the probability of the *explanandum* in any population that is homogeneous with respect to all 'other' possible causes of the *explanandum.* Formally, demand an increase in the *partial conditional probability.*

This move is also suggested by what is called *Simpson's paradox.*[12] Take any fact about the conditional probability of one factor, say A, on another, B. B may increase the probability of A or decrease it or leave it unchanged. Consider a third factor, C, which is probabilistically dependent on both. Then, depending on how the numbers work out, if we stratify on $\pm C$, the original relations between A and B can be shifted in any way at all: B may increase the probability of A in both subpopulations, or decrease it, or leave it unchanged. A standard example is the Berkeley graduate school, which appeared *prima facie* to be discriminating against women: the probability of admission given one was a woman was lower than the probability of admission if one was a man. But, department by department, this turned out not to be the case. What was happening was that women were applying to the departments that were more difficult to get into. The 'true' relation between admission and sex is revealed by stratifying on departments.

The move to demand increase in probability conditional on a fixed arrangement of all other possible causal factors was widely adopted.[13] This brought the standard account of explanation into parallel with Patrick Suppes's probabilistic theory of causality[14], which was published one year before Salmon's influential

[12] Cartwright 1979

[13] Skyrms 1980, Eells 1991, Cartwright 1979

[14] Suppes 1970

paper on probability-lowering explanations. Suppes's probabilistic theory proceeded by two steps. One factor, C, is a *prima facie* cause of a second, E, if C raises the probability of E. A *prima facie* cause is a real cause if and only if C continues to raise the *conditional* probability of E in subpopulations that are homogeneous with respect to all potential confounding factors of E (later authors to these to be all 'other' causes).

One problem for Suppes's account is a problem shared by all accounts that rely on stratification: it cannot deal adequately with cases of purely probabilistic causes where the causes produce their effects in tandem. Consider a purely probabilistic cause that produces a side effect in correlation with an intended effect. In the population that is homogeneous with respect to the presence of the joint cause, the side effect will increase the probability of the effect even though it does not cause it. (This does not happen for deterministic causes because, in the presence of the cause, both the effect and the side effect will have probability one.) Increase in the conditional probability of one factor on another thus cannot be a sufficient condition for the one factor to cause the other in cases where causes may produce their effects probabilistically.

Other problems arise for Suppes from the fact that he takes the two steps separately. Causes may not raise the probability of their effects in a given population for at least two kinds of reasons. First, background correlations may conceal the expected increase in probability, as we can see from the amended version of Salmon's set-up in which the two radioactive elements are equally strong, so the conditional probability of a decay is the same no matter which element is in the box.

Second, a single cause may itself have different capacities with respect to the same effect and these capacities may balance out. Gerhard Hesslow's case of birth control pills and thrombosis is the standard example here.[15] Birth control pills do have a positive capacity to produce thrombosis. They also of course can inhibit pregnancy, and pregnancy can also cause thrombosis. Depending on the strengths of these various capacities, birth control pills may have no overall effect at all on the rate of thrombosis: thrombosis may be equally probable both with and without birth control pills. Cases like this are likely to be common in the social and medical sciences. Consider a cause that produces a deleterious effect. If we want the cause for some other reason, or the cause is difficult or expensive to eliminate, we often try hard to enhance the cause's opposite operations in order to cancel out the unwanted effect. For these kinds of reasons there was a tendency not to adopt Suppes's two-stage definition but rather to move directly to partial conditional probabilities: C explains E just in case C raises the partial conditional probability of E in populations homogeneous with respect to all other factors that can cause E.

By this time the switch-over from explanation to causation was complete. Not only was the account in terms of partial conditional probabilities often explicitly offered as an account of probabilistic causality, but it is difficult to see how it could be thought of in any other way. What after all could defend this

[15] Hesslow 1976, pp. 290-2

particular, very peculiar, choice of probabilistic relation as a mark of *explanation* other than the assumption that the explanations at stake involve *causal* factors?

I have explained that under the influence of Hume and the Positivists philosophers and scientists alike have often been eager to rid science of the concept of causality. One standard strategy has been to 'reduce it away'; that is, to eliminate all use of causal notions and define causation purely in terms of regular association plus, perhaps, some other non-causal concepts, like temporal succession and spatio-temporal contiguity. Nowadays the demand for absolutely regular association has given way. In a move that parallels Hempel's introduction of the I-S model for explanation, purely probabilistic associations are allowed as well. Could we then turn the account I have just described into a reductive definition of causation? The answer is 'No'. The formula involving partial conditional probability cannot serve as a reductive account of causal relevance in terms of probabilistic association because, as it stands, we must condition on other *causal* factors. If we leave out any factors that are causal, we can certainly get mistaken results, as in the example of the Berkeley graduate school. Exactly the same is true if we include too many. Consider a population that is homogeneous with respect to a given disease. We might, for instance, be studying such a population unintentionally, perhaps even without recognizing it, if we are using records from a given hospital. Imagine the disease has two causes, A and B. In that population the absence of A will increase the probability of B, even though not-A does not cause B; and this will be the case even if we focus on a subpopulation that is homogeneous with respect to all the real causes of B.

So, what of the Humean programme of reducing causation away? The formula in terms of partial conditional probabilities has real problems. Many come out immediately when we try to explain what is meant by the expression 'all other causes'. (For instance, in testing C *causes* E, we must not for any particular individual case include any cause of E, say D, that occurs in the causal 'chain' between C and E. That is because, once D is given, information about whether C obtains or not will no longer make any difference to the probability of E. On the other hand, if we do not insist that the test population be homogeneous with respect to D in individual cases where D occurs on its own, not as a part of the process by which C produces E, then Simpson's paradox problems may readily arise.) Despite these real problems, this formula is the best on offer so far if we are looking for some one probabilistic relation between causation and probabilistic association. For Humeans then the question for future work is clear: How do we improve on this formula and make explicit what it means without invoking causal notions?

The admission that we are after all looking for a theory of *causal* explanation provides some progress on another of the central difficulties that the D-N and I-S models were thought to face – the problem of explanatory symmetry. The problem is especially visible for law claims expressed in functional form, where we can rewrite the equations to make any one of the variables appear as the dependent, or *explanandum,* variable. The canonical example is due to Sylvain

Bromberger.[16] When the line of sight of the sun across the top of a flagpole is at an angle θ with the ground, the height of the flagpole (h) and the length of the shadow it casts are related thus: $l = h \sin\theta$. Under the D-N model, we can thus explain why the length takes a given value, l, by citing this law and the height of the pole. That sounds fine. But we can equally use the equation and the length of the flagpole to explain the height of the flagpole, which seems entirely inappropriate in all but the most special of historical situations.

If we turn away from the D-N model, however, and call not for mere deliberately from law, but rather demand that explanations in science cite factors that are causally relevant, the problem does not appear. Causal-law claims come with the asymmetry between cause and effect built in. When the causal-law claims are expressed in the form of equations or, following John Stuart Mill or John Mackie[17] as logical equivalences, the asymmetry is often expressed by writing the effect on the left-hand side and the cause on the right. Nowadays sometimes even a special symbol is inserted to signify both functional equality and causality from right to left.[18]

The identification of explanatory relevance with causal relevance opens a line of attack alternative to the probabilistic accounts growing out of the D-N and I-S models. As a rule of thumb it is fair to say that a controlled experiment is the best way to test a causal claim. Perhaps we can read off a characterisation of causal relevance from an appropriate account of exactly how a controlled experiment should be constructed. This is, for instance, the strategy for characterizing causal/explanatory relevance championed by James Woodward[19] as well as by a number of social scientists. Accounts of this kind are often called *manipulation* or *agency accounts*.[20]

Sometimes manipulation accounts make the strong claim that a causal connection exists only where *we* can manipulate the cause to bring about the effect. This echoes one of Hume's suggestions that our idea of causation could come from our own experience of making things happen in ourselves. As a reductive account of causation, this is a non-starter: the relevant sense of 'manipulation' is that we make the cause vary; but 'making' is itself a causal concept.

But what if we do not add the additional demand that manipulation be done by us? What if we leave out all mention of what makes the cause vary and insist only that it vary in the way necessary for a good experiment? Again, as a reductive account of causation, this is a non-starter. The basic idea of the controlled experiment is to look for variation in the effect as the cause varies by itself; that is, while all other sources of variation in the effect remain constant. Usually, of course, we do not know what all these other possible sources are. The standard strategy in that case is to randomise. We assign

[16] According to Wesley Salmon (*e.g.* 1998, ch. 19), Bromberger never published this example himself. It is, however, widely discussed by philosophers (see *e.g.* van Fraassen 1980, pp. 104-5).

[17] Mackie 1972. For instance, $E = (C_1 + A_1)$ v $(C_2 + A_2)$ ... v $(C_n + A_n)$, where C_i is a salient factor we would normally cite as a cause and A_i is an auxiliary often left unmentioned.

[18] Hoover 2001, Cartwright 2001

[19] Woodward 1997. See also Hausman 1998, Holland 1986, Cartwright forthcoming.

[20] *Cf.* Menzies and Price 1993.

individuals randomly to different values. This makes things even worse for the hopes to read off a reductive account of causality from our best experimental procedures. What makes a good experiment, an experiment able to deliver sound conclusions? In the first place, *all the other causes of variation in the effect* must have the same distribution in each experimental group. In the second, the value that the cause under test takes for each individual in the experiment *must not be caused by any other factors that can influence the effect.* Again, it is difficult to see any way to express these two italicised clauses without the use of causal concepts. Also like the probabilistic account, characterisations in terms of controlled experiments are limited to the kinds of causal systems they can cover and the kinds of causal questions they can answer, as I shall discuss in Section 6. Nevertheless, probabilistic and manipulation accounts are the two primary strategies employed in philosophy of science at this time to characterize general explanatory relevance.

4 The narrowing of the project

The original concern in providing a model of explanation was about explanatory *relevance* of one kind of factor for another. This was clear from Hempel's presentation in his universally used text, *The Philosophy of the Natural Sciences.* As an example, Hempel cites a historical episode.[21] The astronomer Francesco Sizi claimed, against his contemporary Galileo, that there could not be satellites circling around Jupiter and offered the following argument: 'There are seven windows in the head, two nostrils, two ears, two eyes and a mouth; so in the heavens there are two favourable stars, two unpropitious, two luminaries, and Mercury alone undecided and indifferent. From which [...] we gather that the number of planets is necessarily seven'. These features, Hempel complains, do not have the right kind of relation to the planets; they are not the kind of features that *could* explain their arrangement. This echoes the earlier concerns of many of the Positivists about pseudo-sciences. Factors cited in, say, Freudian theory or Marx's theory of history do not have the right kind of internal development nor the right kind of relation to the facts to be explained to count as genuinely explanatory.

This underlying assumption about the point of the project can be seen at each step in the development just described, and it is still there in the probabilistic theory of causality that we now have. The theory is a theory of what *kinds* of factors can cause what other *kinds*, not a theory of what actually causes what on any single occasion. That is, it is not a theory of singular causation, theory of what among all possible causes actually brought about the effect in any given case. This becomes especially apparent in cases of overdetermination. Two factors occur, each of a kind that is causally relevant to the effect. On any such occasion there may be a fact of the matter about which factor produced the effect; but the criteria for determining this are not provided in any of the accounts of explanation described in Section 3.

[21] Hempel 1966, p. 48

This situation is odd because one of the original exemplars of scientific explanantion that Hempel offered was of a case of singular causation: the drop in temperature explains the cracked radiator. Once causation has been readmitted as a legitimate object of study, however, questions about singular causation can be addressed directly in philosophy of science. There are currently a number of tools available for doing so: counterfactuals[22], causal processes[23], agency accounts[24] and single-case probabilities[25].

The gap between explanatory relevance and actual explanation in a single case raises a number of challenges that we must now face:

1. Each of the accounts of singular causation on offer has its own internal problems, as do the accounts of explanatory relevance.

2. There is currently no clear fit between the general accounts of explanatory or causal relevance and accounts of singular causation. For example, must all factors cited in a singular explanation be factors that meet some appropriate criterion for general explanatory or causal relevance in science? Donald Davidson has insisted that the answer is 'yes'.[26] But there are no accounts that have worked out explicit links between the two levels based on either this answer or any other.

3. We have no criteria for adjudicating among the various accounts, either of singular explanation or of general explanatory relevance. Which is best? In Section 6 I will return to this question and offer some proposals about promising directions for future work.

Surprisingly the second major exemplar from Hempel has also been left behind in the developments surrounding explanatory relevance over the last half century. In the original we looked for a model of explanation that could account for how one set of laws can be explained by another more general, more encompassing set; for instance, how Newton's laws explain Kepler's. Accounts that equate explanatory relevance to causal relevance will clearly not do this job. But it is a job that we still want to be able to do. After all, one of the justifications often cited for accepting Newton's laws is that they explain more concrete laws such as Kepler's.[27] This is often called *inference to the best explanation*: we infer that Newton's laws are (probably) true because they provide the best way to explain why Kepler's laws obtain.[28] Is the original D-N model sufficient to handle these kinds of explanations or must better models

[22] Lewis 1973, 1986, 2000

[23] Salmon 1984, Dowe 1992a, 1992b

[24] Woodward 1997, Menzies and Price 1993

[25] Eells 1991, Humphreys 1989

[26] Davidson 1967

[27] Perhaps we should not altogether abandon the idea that explanation even in these cases has something to do with causation since, at least from the time of Copernicus, explanation of laws has often had a causal connotation.

[28] For discussions of how inference to the best explanation works and whether – and why – it is a reasonable form of inference, see Lipton 1991 and references therein.

now be developed? The question becomes especially pressing if we wish to use inference to the best explanation as a reason to believe in the more general, more encompassing theory.

5 Whence explanation: understanding

One might think that explanation had to do with answering questions, relieving puzzles, supplying missing information or providing understanding. This was the point of Sylvain Bromberger's well-known paper on explanations as answers to *why questions*.[29] But it was not the presumption of the work described in Section 2. Explanation in the tradition described there is an objective matter independent of what we know, what we can grasp, or what helps us to understand. This odd feature of the tradition is part of the reason for the claim that the models of explanation on offer served as formal mode surrogates for the material mode concept of causation.

Bas van Fraassen famously rejected the idea that that there is any such objective notion of causation.[30] He stressed instead that explanation is a pragmatic matter, and in particular a correct explanation will be highly sensitive to the form of the question asked. Following the work of Bromberger and van Fraassen there has been a spate of work on why questions and the pragmatics of explanation.[31]

But little of this work is of help with a problem about understanding that still plagues the philosophy of science. Scientists across both the natural and the social sciences regularly construct models that are highly artificial, seriously oversimplified or blatantly false to the situation being modelled. Nevertheless, they claim, the models provide *understanding* of the target phenomena. James Clerk Maxwell provided mechanical models that could account for various electromagnetic properties of the ether. He could not persuade himself fully that any one of these models might literally describe the inner working of the ether. Still he thought that such mechanical models were essential to a proper understanding of electromagnetism.[32] George Fitzgerald did the same, and he even constructed real physical models, not just paper and pencil ones, ensuring that the gears worked properly and the strings would not get tangled up.[33]

Economists regularly produce models that describe not real economies but small 'analogue' economies that bear little resemblance to the real economy of interest. Then they study the analogue economies of their models, they manipulate them, they derive results about them. The results are not accurate predictions about the real economy, even if we restrict attention to a narrow range of intended or targeted results. What is gained from these models? *Understanding*, we are told.

[29] Bromberger 1966
[30] van Fraassen 1980
[31] See for instance Achinstein 1984, Gardenfors 1980, Richardson 1995 and Sintonen 1984.
[32] Cat 2001
[33] Hunt 1991

A different kind of case is in the models of game theory, for instance models that suppose that individuals are playing the prisoner's dilemma game. A lot of theoretical effort has gone into modelling the possible constellations of this one game – one-shot, finitely iterated, indefinitely iterated, with two or N players, rational, boundedly rational, with common knowledge of rationality or not *etc.*[34] Some of these set ups have even been 'tested' empirically.[35] But little detailed work has been done on what social situations these models could plausibly represent.

Also consider, in evolutionary game theory, the model of choice of neighbourhood in which Thomas Schelling assumes that people have a preference to live in a neighbourhood in which the majority of the dwellers are of the same kind as themselves, but without having a preference for a totally segregated environment.[36] However, if everybody chooses the neighbourhood in this way we end up with totally segregated neighbourhoods in the model. The model is supposed to help us understand how segregation is possible without overt discrimination. But how does it do that since what it describes is not a real process that goes on in the way described in any of the real populations for which we might wish to understand neighbourhood segregation?

The social contract is an example from political theory. The story is of people living in a 'state of nature', a state without government. Because of this, it is supposed, each person prey to each other. To avoid the terrible dangers, they all agree to submit to and obey a governing authority that will police them and provide safe social institutions. This is supposed to help us understand something. Presumably not how we actually came to have governments. Perhaps it helps explain the normative force of government – why we ought to obey it. But we never started in a state of nature and agreed to a social contract. So how does the fable of the social contract help us either to understand or to legitimise government and its restrictions?

These are all different cases, and only a tiny sample. Perhaps they will each require a different account of understanding; perhaps we can find a single account that will cover them all. Perhaps many of the claims to understanding are misplaced. But the need for a philosophical treatment is pressing. When does a false model provide understanding? What kind of understanding is involved? And what practically are we able to do with the understanding that we acquire from a blatantly false model? These are crucial questions about an important and widespread scientific practice – and a practice that frequently informs policy – for which we currently have no good answers.

We should recognise that work on these questions may also improve our accounts in other areas. For instance, the standard account of how we use models to understand the world is that we do so by deriving predictions from the models about the target situations in the world; and we do that by looking at all the deductive consequences of the assumptions of the models, where some accounts allow for some corrections or emendations to the deductive consequences based

[34] Kreps *et al.* 1982
[35] Roth 1988, Cooper *et al.* 1996
[36] Schelling 1971

on facts about the specific situation modelled.[37] Mary Morgan, for instance, argues that we gain understanding from *doing things* with models.[38] In a simple economics model we leave the demand curve fixed and we move the supply curve. What happens to the price? Predictions from our simple model about *the kinds of effects* that result from this kind of change may be relatively accurate even if no prediction about the level of the price itself will be.

How do we know what we should and should not do to a model? Morgan argues that an essential ingredient in a model besides the formal structure we usually talk about is a *story*. The story of the model suggests and constrains how we can manipulate the model and what we can conclude from our manipulations. Even in the case where we intend to do nothing more than look at the deductive consequences of a model, something like the story will be necessary – we do not after all expect all the deductive consequences of a simple or blatantly false model to be of use or we would be immediately stopped in our tracks. But exactly what is a story and how are we to judge which stories are acceptable; and what relation does the story have to existing theory? Following up Morgan's ideas will require new work both on the structure of models and on the nature and use of theory.

6 Whence explanation: causation

We have seen a number of accounts of causal-explanatory relevance, and I noted that there are also a variety of different accounts of singular causal explanation on offer. Which is the correct one? I shall suggest that probably they all are – each for a different kind of causal relation.

Begin with Suppes's theory of probabilistic causality. Suppes's definitions form the basis for powerful new methods for causal inference, developed by Wolfgang Spohn[39], by Judea Pearl[40] and by Peter Spirtes, Clark Glymour and Richard Scheines[41], called *Bayes-nets methods.* These methods suppose, as does Suppes, that all causes are *prima facie* causes – *i.e.*, a cause always increases the probability of its effects. Real causes are those that continue to increase the probability of their effects in populations that are homogeneous with respect to an appropriately chosen set of other causes. The methods also assume a *minimality* condition, which asserts that no more causal structure obtains relative to a given probability distribution than what is necessary to guarantee that causes increase the probability of their effects in the two ways presupposed.

The methods provide a way of making causal inferences from facts about probabilities and any antecedent causal knowledge available by exploiting these three constraints. Imagine, for example, a single effect E with two possible causes, C_1 and C_2. There are four possibilities: C_1 is a cause of E and so

[37] Giere 1999
[38] Morgan 1999
[39] Spohn 1980
[40] Pearl 2000
[41] Spirtes *et al.* 1993

is C_2; $C_1$is and C_2 is not; C_2 is and C_1 is not; and neither are causes. C_1 is a genuine cause just in case it raises the probability of E and it continues to do so conditioning on C_2, *if* C_2 is a genuine cause. C_2 is a genuine cause if it raises the probability of E and continues to do so conditioning on C_1, *if* C_1 is a genuine cause. Bayes-nets programs compute all the possible causal arrangements consistent with these assumptions and they do so entirely reliably. Supposing that the three basic assumptions about causation are satisfied, it can be proven that the methods will never yield a false causal conclusion, although in many cases no conclusion at all will be available or, more usually, a disjunction of causal conclusions.

We have seen, however, that Suppes's conditions are violated in a number of cases. What then are we to make of Bayes-nets methods? Are they no good after all? Clearly that conclusion would be a mistake, for the proof referred to shows us exactly what kinds of causal systems the methods can be applied to: those satisfying the three conditions specified.

What do we do when we have a different kind of system? Consider the situation in which we hypothesise that a particular cause behaves like Hesslow's birth control pills – it does not increase the probability of its effect because it operates in different ways both to produce the effect and to prevent it. In this case we try to stop all the other ways the cause can operate and then look for an increase in probability or an increase in level of the effect due to the hypothesised mode of operation. Often we can do this by physically blocking the causal process by which the other operations are carried; often by conditioning on some factor that occurs in each alternative causal process. What about situations in which we suspect that one factor, A, increases the probability of another, B, not because A causes B but because B is a side effect of A from some common probabilistic cause? One standard thing to do in this case is to manipulate A while keeping fixed all the other causes of B. If B does not change in the appropriate ways, we conclude that A is not a cause.

Turn next to manipulation accounts, which are currently our major alternative strategy for characterising causal/explanatory relevance. This strategy is not universally applicable, however, anymore than our best attempts at a probabilistic account. In vast numbers of causal systems, from the human circulatory system to an automobile carburettor, causes are locked together by the design of the system and cannot vary in the requisite ways. Very often a single cause cannot be changed on its own without varying others as well and very often it cannot be changed without interfering with the very process by which it or some other cause produces its effect.

There are also a variety of specific causal hypotheses that we would not test in this way even were the appropriate manipulations available. Consider a claim of the form, "A is a standing condition, just awaiting a trigger to produce B": for example, "The run-down condition of famine victims is a standing condition just awaiting a trigger to produce a fatal disease"; or, "The dryness arising from the drought is a standing condition just awaiting a trigger to produce a forest fire". In this case we do not manipulate A and look for a variation in B. Rather, we keep A fixed and vary the presence or absence of a variety of readily available

appropriate triggers to see if the effect generally occurs when the trigger does.

The lesson to be learned from these brief considerations is that there are different kinds of causal systems and different ways in which a cause can operate within them; and different methods for testing are appropriate for different specific kinds of causal hypotheses relative to different sets of background assumptions about the system in which the putative causal relation is imbedded.

This is the conclusion that we are beginning to draw. For instance, Laurie Paul[42], Judea Pearl[43] and Christopher Hitchcock[44] all describe a variety of different kinds of singular causal relations that can obtain – such as hasteners, delayers, sustainers, contributors. They offer different counterfactual tests for each of the different kinds of relationship.

At the level of general causal relevance, I am a strong advocate of causal diversity. We need, I believe, a background model of the kind of causal system we are dealing with and of the way by which the putative cause is supposed to operate before we can devise a test, or a characterisation, for it. This means that settling matters of causal relevance requires either a lot of antecedent knowledge or a reasonable success at bootstrapping. This makes causal testing difficult, but not impossible. To proceed, however, we need far better accounts of the kinds of causal systems we may encounter and the variety of ways that a cause may operate within them.[45]

We are used to thinking of *causation* as a single monolithic concept: causal relations all have one single feature in common that distinguishes them from mere association; and there is one single canonical mark of that feature. But we have a vast new project in philosophy of science once we recognize that there are a great variety of causal relations and a great variety of causal systems, and each may have its own way of testing. As Christopher Hitchcock advises, "The goal of a philosophical account of causation should not be to capture *the* causal relation, but rather to capture the many ways in which the events of the world can be bound together".[46]

[42] Paul 1988

[43] Pearl 2000

[44] Hitchcock 2003

[45] This need is the basis for the three-year project *Causality: Metaphysics and Methods* now underway at LSE, funded by the British Arts and Humanities Research Board.

[46] Hitchcock 2003

References

Achinstein, Peter (1984), "The Pragmatic Character of Explanation", *PSA* ii: 275-92, repr. in: Ruben 1993: 326-344.

Alexander, Jason and Brian Skyrms (1999), "Bargaining with Neighbors: is Justice Contagious?", *Journal of Philosophy* 96: 588-598.

Bromberger, Sylvain (1966), "Why Questions", in: R. Colodny (ed.): *Mind and Cosmos*. Pittsburgh: Pittsburgh University Press: 86-111.

Carnap, Rudolf (1928), *Der Logische Aufbau der Welt*, Berlin: Weltkreis-Verlag; trans. R. George, *The Logical Structure of the World*, Berkeley, CA: University of California, 1967.

Carnap, Rudolph (1956), "The Methodological Character of Theoretical Terms". In Herbert Feigl and Michaal Scriven (eds): *Minnesota Studies in the Philosophy of Science*. Vol. 1, Minneapolis.

Cartwright, Nancy (1979), "Causal laws & effective strategies", *Noûs* 13: 419-37.

Cartwright, Nancy (2001), *Measuring Causes: Invariance, Modularity and the Causal Markov Condition,* Measurement in Physics and Economics Discussion Paper Series, DP MEAS 10/00, Centre for the Philosophy of Natural and Social Science, London School of Economics.

Cartwright, Nancy (forthcoming), "On Herbert Simon: How to Get Causes from Probabilities", Causality Technical Reports, London School of Economics.

Cat, Jordi (2001), "On Understanding: Maxwell on the Methods of Illustration and Scientific Metaphor", *Studies in the History and Philosophy of Modern Physics* 32: 395-441.

Cooper, Russell, Doughlas V. DeJong and Thomas W. Ross (1996), "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games", *Games and Economic Behavior* 12: 187-218.

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

Dowe, Phil (1992a), "An Empiricist Defence of the Causal Account of Explanation", *International Studies in the Philosophy of Science* 6: 123-128.

Dowe, Phil (1992b), "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", *Philosophy of Science* 59: 195-216.

Eells, Ellery (1991), *Probabilistic Causality*. Cambridge: Cambridge University Press.

Feigl, Herbert and Grover Maxwell (eds) (1962), *Minnesota Studies in the Philosophy of Science Vol. III. Scientific Explanation, Space, and Time*. Minneapolis: University of Minnesota Press.

Gardenfors, Peter (1980), "A Pragmatic Approach to Explanations", *Philosophy of Science* 47: 404-423.

Giere, Ronald (1988), *Explaining Science*. Chicago: University of Chicago Press.

Giere, Ronald (1999), *Science Without Laws*, Chicago: University of Chicago Press.

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

Hempel, Carl G. (1942), "The Function of General Laws in History", *Journal of Philosophy* 39: 35-48, repr. in Hempel 1965: 231-43.

Hempel, Carl G. (1962), "Deductive-Nomological Explanation vs. Statistical Explanation", in: Feigl et al. 1962, 98-169.

Hempel, Carl G. (1965), *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: Free Press.

Hempel, Carl G. (1966), *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice Hall.

Hempel, Carl G. and Paul Oppenheim (1948), "Studies in the Logic of Explanation", *Philosophy of Science* 15: 135-75, repr. in Hempel 1965: 245-295.

Hesslow, Gerhard (1976), "Discussion: Two Notes on the Probabilistic Approach to Causality", *Philosophy of Science* 43: 290-2.

Hitchcock, Christopher (2003), "Of Humean Bondage", *British Journal for the Philosophy of Science* 54: 1-25.

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

Hoover, Kevin D. (2001), *Causality in Macroeconomics,* Cambridge; Cambridge University Press.

Hume, David (1997/1748): *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. Oxford: OUP.

Humphreys, Paul (1989), *The Chances of Explanation*. Princeton: Princeton University Press.

Hunt, Bruce J. (1991), *The Maxwellians*. Ithaca: Cornell University Press.

Jeffrey, Richard (1971), "Statistical Explanation vs. Statistical Inference", in Salmon *et al.*, 19-28.

Kreps, D., P. Milgrom, J. Roberts and R. Wilson (1982), "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma", *Journal of Economic Theory* 17: 245-252.

Lakatos, Imre (1970), "Falsification and the Methodology of Scientific Research Programmes". In: Imre Lakatos and Alan Musgrave (eds): *Criticism and the Growth of Knowledge*. Cambridge: CUP, pp. 91-196.

Laudan, Larry (1996), *Beyond Positivism and Relativism. Theory, Method and Evidence*. Boulder and Oxford: Westview.

Lewis, David (1973), "Causation", *Journal of Philosophy* 70: 556-67.

Lewis, David (1986), "Postscript to 'Causation"', in his *Philosophical Papers*, Vol. II, Oxford: Oxford University Press, 173-213.

Lewis, David (2000), "Causation as Influence", *Journal of Philosophy* 97: 182-197.

Lipton, Peter (1991), *Inference to the Best Explanation*. London: Routledge.

Mackie, John L. (1972), *The Cement of the Universe*. Oxford: Clarendon Press.

Menzies, Peter and Huw Price (1993): "Causation as a Secondary Quality", *British Journal for the Philosophy of Science* 44: 187-203.

Morgan, Mary S. (1999), "Learning from Models", in Morgan *et al.* (eds): 347-88.

Morgan, Mary S. and Margaret Morrison (eds) (1999), *Models as Mediators*. Cambridge: Cambridge University Press.

Paul, Laurie A. (1998), "Keeping Track of the Time: Emending the Counterfactual Account of Causation", *Analysis* 58: 191-98.

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

Richardson, Alan (1995), "Explanation: Pragmatics and Asymmetry", *Philosophical Studies* 80: 109-29.

Roth, Alvin E. (1988), "Laboratory Experimentation in Economics: A Methodological Overview", *The Economic Journal* 98: 974-1031.

Ruben, David (ed.) (1993), *Explanation*, Oxford: Oxford University Press.

Salmon, Wesley (1971), "Statistical Explanation", in Salmon *et al.*, 29-88.

Salmon, Wesley (1984): *Scientific Explanation and the Causal Structure of the World.* Princeton: Princeton University Press.

Salmon, Wesley (1998), *Causality and Explanation*, Oxford: OUP.

Salmon, Wesley, Richard Jeffrey and James Greeno (eds) (1971), *Statistical Explanation and Statistical Relevance.* Pittsburgh: Pittsburgh University Press.

Skyrms, Brian (1980), *Causal Necessity*, New Haven: Yale University Press.

Skyrms, Brian (1996), *Evolution of the Social Contract,* Cambridge: Cambridge University Press.

Schelling, Thomas C. (1971), "Dynamic Models of Segregation", *Journal of Mathematical Sociology* 1: 143-186.

Scriven, Michael (1959), "Explanation and Prediction in Evolutionary Theory", *Science* 130: 477-82.

Sintonen, Matti (1984), *The Pragmatics of Scientific Explanation*, Acta Philosophica Fennica, Vol. 37, Helsinki; Academic Bookstore.

Spirtes, Peter, Clark Glymour and Richard Scheines (1993), *Causation, Prediction and Search,* New York: Springer.

Spohn, Wolfgang (1980), "Stochastic Independence, Causal Independence, and Shieldability", *Journal of Philosophical Logic*, 9: 73-99.

Suppe, Frederick (1989), *The Semantic Conception of Theories and Scientific Realism*, Chicago: University of Illinois Press .

Suppes, Patrick (1967): "What Is a Scientific Theory?" In: Sidney Morgenbesser (ed.): *Philosophy of Science Today.* New York: Basic Books.

Suppes, Patrick (1970), *A Probabilistic Theory of Causality*, Amsterdam: North-Holland.

van Fraassen, Bas (1980), *The Scientific Image*. Oxford: Clarendon Press.

Woodward, James (1997), "Explanation, Invariance, and Intervention", *PSA* 1996, ii: S26-41.