Chapter 3

Explanation

Jim Woodward

Although the subject of explanation has been a major concern of philosophy since Plato and Aristotle, modern philosophical discussion of this topic, at least as it pertains to science, begins with the so-called deductive-nomological (DN) model of explanation in the middle of the twentieth century. This model has many advocates but unquestionably the most detailed and influential statement is due to Carl Hempel (1965).

The DN Model

The basic idea of the DN model is that explanations have the structure of sound deductive arguments in which a law of nature occurs as an essential premise. One deduces the *explanandum*, which describes the phenomenon to be explained, from an *explanans*, consisting of one or more laws, typically supplemented by true sentences about initial conditions. The model is intended to apply both to the explanation of "general regularities" by other laws and the explanation of particular events, although subsequent developments have largely focused on the latter. The derivation of facts about planetary trajectories (e.g. Kepler's laws) from the laws of Newtonian mechanics, the gravitational inverse square law and appropriate information about initial conditions is a paradigmatic illustration of the pattern of explanation that the DN model attempts to capture.

The DN model is meant to capture explanation via deduction from deterministic laws and this raises the obvious question of the explanatory status of statistical laws. Hempel claims that there is a distinctive sort of statistical explanation, which he calls inductive-statistical or IS explanation, involving the subsumption of individual events (like the recovery of a particular person from streptococcus infection) under statistical laws (such as a law specifying the probability of recovery, given that penicillin has been taken). The details of Hempel's

account are complex, but the underlying idea is roughly this: an IS explanation will be good to the extent that its explanans confers high probability on its explanandum. Although once a flourishing area of research, the structure of statistical explanation has received relatively little attention recently.[1] In what follows, I will largely ignore it.

Much of the appeal of the DN model lies in the undeniable fact that in some areas of science, such as physics, many explanations do seem to involve derivations from laws. However, the DN model (or at least the version of the model I will discuss) is committed to a good deal more than this commonplace observation. It claims that *all* explanations conform to the requirements of the model, and that everything conforming to those requirements is an explanation. We need to ask whether these claims are correct and whether the key components of the model such as the notion of a law, are sufficiently clear and well-understood to play the role the model assigns to them. I begin with this second issue and then turn to whether the DN requirements are necessary and sufficient for explanation.

Laws

There is general agreement among defenders of the DN approach that laws are (at least) regularities or uniformities – they tell us that if a system exhibits certain properties, it will always or with a certain probability exhibit others. However, not all regularities – even exceptionless regularities – are laws. To take a stock example, while "all spheres of uranium have a mass of less than 10^5 kg" is regarded as a law (since the critical mass for uranium is only a few kilograms), the syntactically similar generalization, "all spheres of gold have a mass of less than 10^5 kg," although presumably true is no law and hence cannot play the role of nomological premise in a DN explanation. The problem of distinguishing genuine laws from such "accidental regularities" is thus central to a defense of the DN model.

Most philosophers, including both defenders and critics of the DN model, have assumed that an adequate account of laws must satisfy certain "empiricist" strictures. These are rarely explained with any precision, but amount in practice to the requirement that the account be "reductive": notions like "law," "cause," and "explanation" are seen as belonging to a family of closely interrelated concepts that must, on pain of "circularity," be explicated in terms of concepts that lie outside of this family like "regularity." A number of criteria for lawfulness that are thought to meet these strictures have been proposed: laws are said

1 to be exceptionless generalizations
2 to contain only purely qualitative predicates and make no reference to particular objects or spatio-temporal locations
3 to support counterfactuals

4 to be confirmable by a limited number of instances in a way that accidental generalizations are not, and
5 to be integrated into some body of systematic theory and play a unifying role in inquiry in a way that accidental generalizations do not.

While each set of criteria has its defenders, I think that a fair summary of current discussion is that none, either singly or in combination, is generally accepted. Many, perhaps most, paradigmatic laws violate certain of the criteria such as (1). Others, such as (2) seem both unclear and overly restrictive and have been abandoned in most recent discussions. Criteria (3) and (5) are, as formulated, both vague and arguably satisfied by accidental as well as lawful generalizations.[2] Criterion (4) looks fundamentally confused from the perspective of any modern treatment of confirmation.[3]

Given the absence of a satisfactory account of lawhood, it is natural to wonder whether the contrast between laws and non-laws can play the central role it is assigned in the DN model. If we cannot say what laws are, why should we accept the DN claim that they are required for successful explanation? One possible response is that although there may be no generally accepted account of laws, there is at least general agreement about which generalizations count as laws and this is all the DN model requires. In fact, however, there seems to be no such agreement. The so-called special sciences – biology, psychology, economics and so on – are full of generalizations that appear to play an explanatory role and/or to describe causal relationships and yet fail to satisfy many of the standard criteria for lawfulness. For example, although Mendel's law of segregation (M) is widely used in evolutionary models, it has a number of exceptions, such as meiotic drive. Other widely used generalizations in the special sciences have very narrow scope in comparison with paradigmatic laws, hold only over restricted spatio-temporal regions, and lack explicit theoretical integration. There is considerable disagreement over whether such generalizations are laws. Some philosophers suggest that such generalizations satisfy too few of the standard criteria to count as laws but can nevertheless figure in explanations; hence we should abandon the DN requirement that all explanations must appeal to laws. Others – e.g. Mitchell (1997) – emphasizing different criteria for lawfulness, conclude instead that generalizations like (M) are laws and hence no threat to the requirement that explanations invoke laws. In the absence of an adequate account of laws, it is hard to evaluate these competing claims.

Motivation

Putting aside these unclarities surrounding the notion of law, why suppose that all (or even some) explanations have a DN or IS structure? Hempel appeals to two central motivating ideas. The first connects the information provided by a DN argument with a certain conception of what it is to achieve understanding:

> a DN explanation answers the question "Why did the explanandum-phenomenon occur?" by showing that the phenomenon resulted from certain particular circumstances, specified in $C_1, C_2, \ldots, C_k$, in accordance with the laws $L_1, L_2, \ldots, L_r$. By pointing this out, the argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon was to be expected; and it is in this sense that the explanation enables us to understand why the phenomenon occurred (Hempel, 1965, p. 337).

IS explanation involves a natural generalization of this idea: it shows that the explanandum-phenomenon was to be expected, on the basis of a law, with high probability.

The second main motivation for the DN/IS (hereafter DN) model has to do with the role of causation in explanation. Whether or not all explanations are causal – itself a disputed question in the theory of explanation – there is general agreement among philosophers that many explanations cite information about causes. However, most philosophers, including advocates of the DN model like Hempel, have been unwilling to take the notion of causation as primitive in the theory of explanation. Instead, they have regarded the notion of causation as at least as much in need of explication as the notion of explanation and have sought an account of causation meeting the reductionist or empiricist requirements described above in connection with notion of law. While there are many forms that a theory of causation might take, advocates of the DN model have generally accepted a broadly Humean or regularity theory of causation, according to which (very roughly) all causal claims imply the existence of some corresponding law or regularity linking cause to effect. This is then taken to show that all causal explanations "imply," perhaps only "implicitly," the existence of some law and hence that laws are "involved" in all such explanations, just as the DN model claims.

To illustrate of this line of argument, consider

> (Ex1) The impact of my knee on the desk caused the tipping over of the inkwell.

(Ex1) is a so-called singular causal explanation, advanced by Michael Scriven (1962) as a counterexample to the claim that the DN model describes necessary conditions for successful explanation. According to Scriven, (Ex1) explains the tipping over of the inkwell even though no law or generalization figures explicitly in (Ex1) and (Ex1) appears to consist of a single sentence, rather than a deductive argument. Hempel's response (1965, p. 360) was that (Ex1) should be understood claiming there is a "law" or regularity linking knee impacts to tipping over of inkwells. It is the claim that some such law holds that "distinguishes" (Ex1) from "a mere sequential narrative" in which the spilling is said to follow the impact but without any claim of causal connection. We should think of this law as the nomological premise in the DN argument that, according to Hempel, is "implicitly" asserted by (Ex1). Critics have in turn responded that the claim that (Ex1) implies, in virtue of its meaning, the existence of an underlying DN argu-

ment looks implausible, given the fact that people use and understand such explanations even if they lack the concepts like "deductively valid argument" and "law of nature."

Counterexamples

While (Ex1) is a potential counterexample to the claim that the DN model provides necessary conditions for explanation, several other examples challenge the claim that the DN model provides sufficient conditions.

Many explanations exhibit directional or asymmetric features to which the DN model appears to be insensitive. From information about the height (h) of a flag pole, the angle ϕ it makes with the sun, and laws describing the rectilinear propagation of light one can deduce the length (s) of its shadow – such a derivation is arguably an explanation (call it (Ex2)) of s. It is equally true that from s, these same laws, and ϕ, one can deduce h. Such a derivation (Ex3) although it apparently meets all of the criteria for an acceptable DN argument, is no explanation of why the flagpole has this height (Bromberger, 1966).

There are other kinds of explanatory irrelevancies besides those associated with the directional features of explanation. Consider a well-known example due to Wesley Salmon (1971).

(Ex4) (L) All males who take birth control pills regularly fail to get pregnant.
John Jones is a male who has been taking birth control pills regularly.
John Jones fails to get pregnant.

(L) appears to meet the criteria for lawfulness accepted by Hempel and many other writers.[4] Despite this, (Ex4) is no explanation of why Jones fails to get pregnant.

Since both of these derivations show that their putative explananda were "nomically expectable," they seem to cast doubt on the whole idea that explaining an outcome is (just) a matter of showing that it was to be expected on the basis of a law.

One obvious diagnosis of both examples is that they neglect the role that causation plays in explanation. The height of the flagpole causes the length of its shadow and this is why we find a derivation of the former from the latter explanatory. By contrast, the length of the shadow is an effect, not a cause of the height of the flagpole and this is why we don't regard a derivation of h from s as explanatory. Similarly, taking birth control pills does not cause Jones' failure to get pregnant and this is why (Ex4) is not an acceptable explanation.

As explained above, advocates of the DN model would not regard this diagnosis as very illuminating, unless accompanied by some positive account of causation. We should note, however, that an apparent lesson of (Ex3) and (Ex4) is that the regularity account of causation favored by DN theorists is at best incom-

plete: the occurrence of *c*, *e* and the existence of some law linking them (or *x's* having property *P* and *x's* having property *Q* and some law linking these) is at best a necessary and not a sufficient condition for the truth of the claim that *c* caused *e* or *x's* having *P* is causally or explanatorily relevant to *x's* having *Q*. Contrary to what is often claimed – see, for example Kim (1999, p. 17) – we can not argue that explanations like (Ex1) have an implicit DN structure on the grounds that instaniations of such a structure "guarantee" that *c* is causally or explanatorily relevant to *e*.

The SR Model

To a significant extent, subsequent developments in the theory of explanation represent attempts to capture the features of causal or explanatory relevance that appear to be left out of examples like (Ex3) and (Ex4), usually within the empiricist constraints described above. Wesley Salmon's statistical relevance (or SR) model (Salmon, 1971) attempts to capture these features in terms of the notion of statistical relevance (conditional dependence relationships). On the SR model, a request for explanation will take the following canonical form: Why does this member *x* of the class characterized by attribute *A* have attribute *B*? Define a *homogenous partition* of *A* as a set of subclasses or cells C_i of A that are mutually exclusive and exhaustive, where $P(B|A.C_i) \neq P(B|A.C_j)$ for all $C_i \neq C_j$ and where no further statistically relevant partition of any of the cells $A.C_i$ can be made with respect to *B* – that is, there are no additional attributes D_k in *A* such that $P(B|A.\ C_i) \neq P(B|A.\ C_i.\ D_k)$. Then an SR explanation of why *A* is *B* consists of

(i) the prior probability of *B* within *A*: $P(B|A) = p$
(ii) a homogeneous partition of *A* with respect to *B*, $(A.\ C_1, \ldots A.C_n)$, together with the probability of *B* within each cell of the partition: $P(B|A.Ci) = p_i$, and
(iii) The cell of the partition to which *x* belongs.

To employ one of Salmon's examples, suppose we want to construct an SR explanation of why *x* who is a teenager (= *A*) is delinquent (= *B*). Suppose further that there just two attributes and no others that are statistically relevant to *B* in *A* – gender (*M* or *F*) and whether residence is urban (*U*) or rural (*R*), with the probability of *B* conditional on *A* and each the four possible conjunctions of these attributes being different. Then {*A.M.U, A.M.R, A.F.U, A.F.R*} is a homogenous partition of *A* with respect to *B* and the SR explanation will consist of

(i) a statement of the probability of being a delinquent within the class of teenagers

(ii) a statement of the probability of delinquency within this class as we condition on each of the four possible combinations of attributes, and
(iii) the cell to which x belongs.

Intuitively, the idea is that this information tells us about the relevance of each of these combinations of attributes to being delinquent among teenagers and has explanatory import for just this reason. As an additional illustration, suppose that in the birth control pills example (Ex4) the original population T includes both genders. Then

$$\mathrm{P}(\textit{Pregnancy}|T.\ \textit{Male}.\ \textit{Takes birth control pills}) = \mathrm{P}(\textit{Pregnancy}|T.\ \textit{Male})$$

while

$$\mathrm{P}(\textit{Pregnancy}|T.\ \textit{Male}.\ \textit{Takes birth control pills}) \neq \mathrm{P}(\textit{Pregnancy}|T.\ \textit{Takes birth control pills})$$

assuming that birth control pills are not always effective for women. In this way, we can capture the idea that among males, taking birth control pills is explanatorily irrelevant to pregnancy, while being male *is* relevant.

The SR model has a number of features that have generated substantial discussion, but I want to focus on what I take to be the central motivating ideas of the model:

(i) explanations cite causal relationships and
(ii) causal relationships are captured by statistical relevance relationships.

The fundamental problem with the SR model is that (ii) is false – as a substantial body of work[5] has made clear, casual relationships are greatly underdetermined by statistical relevance relationships. Consider Salmon's example of a system in which atmospheric pressure A is a common cause of the occurrence of a storm S and the reading of a barometer B with no causal relationship between B and S. Salmon claims that B is statistically irrelevant to S given A – i.e. $\mathrm{P}(S|A.B) = \mathrm{P}(S|A)$ but A remains relevant to S given B – i.e. $\mathrm{P}(S|A.B) \neq \mathrm{P}(S|B)$ and thus that A is explanatorily (causally) relevant to S while B is not. However, many other causal structures are compatible with these statistical relevance relationships. Structures in which B causes A which in turn causes S will, if we make assumptions like Salmon's connecting causation and probability, lead to exactly the same statistical relevance relationships. In these structures, unlike Salmon's example, B is causally (and presumably explanatorily) relevant to S. Similarly, the statistical relevance relationships among A, B and S, will not tell us whether we are dealing with a system in which, say, A causes B which causes S and in which A also directly causes S, independently of B, or one in which the direction of the causal arrow from A to B is reversed, so that B causes A. A mere list of statistical relevance relationships, which is what the SR model provides, does not tells us which causal or explanatory relationships are operative.

The Causal Mechanical Model

In more recent work, Salmon (1984) acknowledges this and abandons the attempt to characterize explanation or causal relationships in purely statistical terms. His new account, which he calls the Causal Mechanical (CM), attempts to capture the "something more" involved in causal/explanatory relationships over and above facts about statistical relevance. The CM model employs several central ideas. A *causal process* is a physical process, like the movement of a particle through space, that is characterized by the ability to transmit its own structure in a continuous way. A distinguishing feature of causal processes is their ability to transmit marks. Intuitively a mark is some local modification to the structure of a process, as when one scuffs the surface of a baseball. A baseball is a causal process and one expects the scuff mark to persist as the baseball moves from one spatio-temporal location to another, even in the absence of further interventions or interactions. Causal processes contrast with *pseudo-processes* which lack the ability to transmit marks. An example is the shadow of a moving physical object. Intuitively, Salmon's idea is that, if we try to mark the shadow by modifying its shape at one point (for example, by altering a light source or introducing a second occluding object), this modification will not persist unless we continually intervene to maintain it as the shadow occupies successive spatio-temporal positions. *Causal interactions* occur when one causal process spatio-temporally intersects another and produces a modification of it structure. An example would be a collision between two particles which alters the direction and kinetic energy of both.

According to the CM model, an explanation of some event *E* will trace the causal processes and interactions leading up to *E* (Salmon calls this the *etiological* aspect of the explanation), or at least some portion of these, as well as describing the processes and interactions that make up the event itself (the *constitutive* aspect of explanation). In this way, the explanation shows how *E* "fit[s] into a causal nexus" (1984, p. 9).

The suggestion that explanation involves "fitting" an explanandum into a causal nexus does not of course give us any very precise characterization of just what the relationship between *E* and other causal processes and interactions must be if information about the latter is to explain *E*. But rather than belaboring this point, I will focus on the intuitive idea behind this suggestion and examine what implies for some specific examples.

Suppose that a cue ball, set in motion by the impact of a cue stick, strikes a stationary eight ball with the result that the eight ball is put in motion and the cue ball changes direction. The impact of the stick also transmits some blue chalk to the cue ball which is then transferred to the eight ball on impact. The cue stick, the cue ball and the eight ball are causal processes and the collision of the cue stick with the cue ball and the collision of the cue and eight balls are causal interactions. Salmon's intuitive idea is that citing such facts about processes and interactions explains the motion of the balls after the collision; by contrast, if one

of these balls casts a shadow that moves across the other, this will be causally and explanatorily irrelevant to its subsequent motion since the shadow is a pseudo-process.

However, as Christopher Hitchcock shows in an illuminating paper (Hitchcock, 1995) the information about causal processes and interactions just described leaves out something important. The usual elementary textbook "scientific explanation" of the motion of the balls following collision proceeds by deriving that motion from information about their masses and velocity before the collision, the assumption that the collision is perfectly elastic, and the law of the conservation of linear momentum. We think of the information conveyed by this derivation as showing that it is the mass and velocity of the balls, rather than, say, their color or the presence of the blue chalk mark, that is explanatorily relevant to their subsequent motion. However, it is hard to see what in the CM model allows us to pick out the linear momentum of the balls, as opposed to various other features, as explanatorily relevant. Part of the difficulty is that to express such relatively fine-grained judgments of explanatory relevance (that it is linear momentum rather than chalk marks that matter) we need to talk about relationships between properties or magnitudes and it is not clear how express such judgments in terms of facts about causal processes and interactions. Both the linear momentum and the blue chalk mark communicated to the cue ball by the cue stick are marks that are transmitted by the spatio-temporally continuous causal process consisting of the motion of the cue ball, and which then are transmitted via an interaction to the eight ball.

Ironically, as Hitchcock goes on to note, a similar observation may be made about (Ex4). Spatiotemporally continuous causal processes that transmit marks as well as causal interactions are at work when male Mr. Jones ingests birth control pills – the pills dissolve, components enter his bloodstream, are metabolized or processed in some way and so on. Similarly, causal processes (albeit different processes) and spatio-temporally continuous paths are at work when female Ms. Jones takes birth control pills. Intuitively, it looks as though the relevance or irrelevance of the birth control pills does not just have to do with whether the actual processes that lead up to Mr. Jones non-pregnancy are capable of mark transmission but rather (roughly) with the contrast between what happens in actual situation in which Jones takes the pills and an alternative situation in which Jones does not take the pills. It is because the outcome (non-pregnancy) would be the same in both cases if Jones is male that the pills are explanatorily irrelevant. This links explanatory relevance to counterfactuals – a point to which I will return.

A second, not unrelated set of worries has to do with how we are to apply the CM model to more complex systems which involve a large number of interactions among what from a fine grained level of analysis are distinct causal processes. Suppose that we have a mole of gas, confined to a container, with volume V_1, at pressure P_1, and temperature T_1. The gas is then allowed to expand isothermally into a larger container of volume V_2. One standard way of explaining the behavior of the gas – its rate of diffusion and its subsequent equilibrium pressure P_2 –

appeals to the generalizations of phenomenological thermodynamics – e.g., the ideal gas law, Graham's law of diffusion, etc. Salmon appears to regard putative explanations based on at least the first of these generalizations as not really explanatory because they do not trace continuous causal processes – the individual molecules are causal processes but not the gas as a whole. However, it is obviously impossible to trace the causal processes and interactions represented by each of the 6×10^{23} molecules making up the gas and the successive interactions (collisions) it undergoes with every other molecule. The usual statistical mechanical treatment, which Salmon presumably would regard as explanatory, does not attempt to do this. Instead, it makes certain general assumptions about the distribution of molecular velocities and the forces involved in molecular collisions and then uses these, in conjunction with the laws of mechanics, to derive and solve a differential equation (the Boltzmann transport equation) describing the overall behavior of the gas. This treatment abstracts radically from the details of the causal processes involving particular individual molecules and instead focuses on identifying higher level variables that aggregate over many individual causal processes and that figure in general patterns that govern the behavior of the gas. A plausible version of the causal mechanical model will need to avoid the conclusion that an explanation of the behavior of the gas must trace the trajectories of individual molecules and provide an alternative account of what tracing causal processes and interactions means for such a system. Such an extension of the CM model has not yet been developed. A similar point holds for other complex systems.[6]

There is another aspect of this example that is worthy of comment. Even if, per impossible, an account that traced individual molecular trajectories were to be produced, there are important respects in which it would not provide the explanation of the macroscopic behavior of the gas that we are looking for. This is because there are a very large number of different possible trajectories of the individual molecules in addition to the trajectories actually taken that would produce the macroscopic outcome that we want to explain. Very roughly, given the laws governing molecular collisions one can show that almost all (i.e., all except a set of measure zero) of the possible initial positions and momenta consistent with the initial macroscopic state of the gas, as characterized by P_1, T_1, and V_1, will lead to molecular trajectories such that the gas will evolve to the macroscopic outcome in which the gas diffuses to an equilibrium state of uniform density through the chamber at pressure P_2. Similarly, there is a large range of different microstates of the gas compatible with each of the various other possible values for the temperature of the gas and each of these states will lead to a different final pressure P_2*. It is an important limitation of the strategy of tracing actual individual molecular trajectories that it does not, at least as it stands, capture or represent this information. Explaining the final pressure P_2 of the gas seems to require identifying both the full range of (counterfactual and not just actual obtaining) conditions under which P_2 would have occurred and the (counterfactual) conditions under which it would have been different. Just tracing the causal processes (in the form of actual molecular trajectories) that lead to P_2, as the CM model

requires, omits this information about what would happen under these counterfactual conditions.

Unificationist Models

The final account of explanation that we will examine is the *unificationist* account. The basic idea was introduced by Michael Friedman (1974) but its subsequent development has been most associated closely with Philip Kitcher (1989). One possible assessment of the DN model is that it (or something broadly like it) is correct as far as it goes – it states plausible necessary conditions on explanation – but that it needs to be supplemented by some additional condition *X* which avoids the counterexamples to the sufficiency of the model described above. This is roughly Kitcher's view. Explanations are derivations from premises that include generalizations of considerable scope (whether or not we regard these as laws) but such derivations must also meet an additional condition = *X* having to do with unification. The underlying idea is that explanatory theories are those that unify a range of different phenomena. Such unifications clearly have played an important role in science; paradigmatic examples include Newton's unification of terrestrial and celestial theories of motion and Maxwell's unification of electricity and magnetism.

Kitcher attempts to make this idea more precise by suggesting that explanation is a matter of deriving as many descriptions as possible of different phenomena by using the same "argument patterns" over and over again – the fewer the patterns used, the more "stringent" they are in the sense of imposing restrictions on the derivations that instaniate them, and the greater the range of different conclusions derived, the more unified our explanations. Kitcher does not propose a completely general theory of how these considerations – number of conclusions, number of patterns, and stringency of patterns – are to be traded off against one another, but he does suggest that, in many specific cases, it will be clear enough what these considerations imply about the evaluation of particular candidate explanations. His basic strategy is to argue that the derivations we regard as good explanations are instances of patterns that taken together score better according to the criteria just described than the patterns instantiated by the derivations we regard as defective explanations. Following Kitcher, let us define the *explanatory store E(K)* as the set of argument patterns that maximally unifies *K*, the set of beliefs accepted at a particular time in science. Showing that a particular derivation is an acceptable explanation is then a matter of showing that it belongs to the explanatory store.

As an illustration, consider Kitcher's treatment of the problem of explanatory asymmetries. Our present explanatory practices – call these *P* – are committed to the idea that derivations of a flagpole's height from the length of its shadow are not explanatory. Kitcher contrasts *P* with an alternative systemization in which such derivations are regarded as explanatory. According to Kitcher, *P* includes the

use of a single origin and development (OD) pattern of explanation, according to which the dimensions of objects – artifacts, mountains, stars, organisms etc. – are traced to "the conditions under which the object originated and the modifications it has subsequently undergone" (1989, p. 485). Now consider the consequences of adding to *P*, an additional pattern *S* (the shadow pattern) which permits the derivation of the dimensions of objects from facts about their shadows. Since the OD pattern already permits the derivation of all facts about the dimensions of objects, the addition of *S* to *P* will increase the number of argument patterns in *P* and will not allow us to derive any new conclusions. On the other hand, if we were to drop OD from *P* and replace it with the shadow pattern, we would have no net change in the number of patterns in *P* but would be able to derive far fewer conclusions than we would with OD, since many objects do not have shadows from which to derive their dimensions. Thus OD belongs to the explanatory store, and the shadow pattern does not. Kitcher's treatment of other problem cases in the theory of explanation is similar – for example, derivations like (Ex4) above are claimed to instantiate patterns that belong to a totality of patterns that are less unifying than the totality to which the pattern instantiated by a derivation that just appeals to a generalization about all males failing to become pregnant.

What is the role of causation on this account? Kitcher claims that "the 'because' of causation is always derivative from the 'because' of explanation" (1989, p. 477). That is, our causal judgments simply reflect the explanatory relationships that fall out of our (or our intellectual ancestors') attempts to construct unified theories of nature. There is no independent causal order over and above this which our explanations must capture.

Although the idea that explanation has something to do with unification is intuitively appealing, Kitcher's particular way of cashing out the idea seems problematic. His treatment of the flagpole example obviously depends heavily on the contingent truth that some objects do not cast shadows. But wouldn't it still be inappropriate to appeal to facts about the shadows cast by objects to explain their dimensions in a world in which all objects cast enough shadows (they are illuminated from a variety of different directions etc.) so that all of their dimensions can be recovered?[7]

The matter becomes clearer if we turn our attention to a variant example in which, unlike the shadow example, there are clearly just as many backwards derivations from effects to causes as there are derivations from causes to effects. Consider, following Barnes (1992), a time-symmetric theory like Newtonian mechanics, as applied to a closed system like the solar system. Call derivations of the state of motion of the particles at some future time t from information about their present positions (at time t_0), masses, and velocities, the forces incident on them between t_0, and the laws of mechanics *predictive*. Now contrast such derivations with *retrodictive* derivations in which the present motions of the particles are derived from information about their future velocities and positions at t, the forces operative between t_0 and t and so on. It looks as though there will be just as many retrodictive derivations as predictive derivations and each will require premises of

exactly the same general sort – information about positions, velocities, masses etc. and the same laws. Thus, the pattern or patterns instantiated by the retrodictive derivations looks exactly as unified as the pattern or patterns associated with the predictive derivations. However, we think of the predictive derivations and not the retrodictive derivations as explanatory and the present state of the particles as the cause of their future state and not vice-versa. It is far from obvious how considerations having to do with unification could generate such an explanatory asymmetry.

Examples of this sort cast doubt on Kitcher's claim that one can begin with the notion of explanatory unification, understood in a way that does not presuppose causal notions, and use it to derive the content of causal judgments. This conclusion is reinforced by a more general consideration: The conception of unification underlying Kitcher's account is, at bottom, one of descriptive economy or information compression – deriving as much from as few assumptions or via as few patterns of inference as possible. However, there are many schemes and procedures in science that involve information compression and unified description but don't seem to provide information about causal relationships. This is true of many classificatory schemes including schemes for biological classification, and schemes for the classification of geological and astronomical objects like rocks and stars. If I know that individuals belong to a certain classificatory category (e.g. *X*s are mammals), I can use this information to derive a great many of their other properties (*X*s have backbones, hearts, their young are born alive, etc.) and this is a pattern of inference that can be used repeatedly for many different sorts of *X*s. Nonetheless, and despite the willingness of some philosophers to regard such derivations as explanatory (*X* is white because *X* is a polar bear and all polar bears are white), most scientists think of such schemes as "merely descriptive" and as telling us little or nothing about the causes or mechanisms that explain why *X*s have hearts or are white. Similarly, there are numerous statistical procedures (factor analysis, cluster analysis, multi-dimensional scaling techniques) that allow one to summarize or represent large bodies of statistical information in an economical, unified way and to derive more specific statistical facts from a much smaller set of assumptions by repeated use of the same pattern of argument. For example, knowing the "loading" of each of n intelligence tests on a single common factor *g*, one can derive $n(n-1)/2$ conclusions about pairwise correlations among these tests. Again, however, it is doubtful that this "unification" tells us anything about causal relationships.

Conclusion and Directions for Future Work

What conclusions/morals may we draw from this historical sketch? What are the most promising directions for future work? Any proposals about these matters will be tendentious, but with this caveat in mind, I suggest the following. First, many

of the limitations of the theories reviewed above may be traced to their failure to satisfactorily capture causal notions. A more adequate account of causation is thus one of the most important items on the agenda for future work on explanation. The approach I regard as most promising differs from those described above – it takes counterfactual dependence to be the key to understanding causation and hence explanation. To motivate this approach, note that an obvious diagnosis of the difference between the acceptable and defective explanations described above is that the former but not the latter exhibit a pattern of counterfactual dependence between explanans and explanandum in the following sense: in the good explanations but not the bad ones, changing the explanans variables will be associated with a corresponding change in the explanandum. Thus, the birth control pills are causally and explanatorily irrelevant to Mr. Jones' pregnancy because whether he becomes pregnant does not depend counterfactually on whether he takes pills. We might establish this absence of counterfactual dependence by doing an experiment in which we observe that manipulating whether males take birth control pills is associated with no change in whether they become pregnant. Similarly, if we change the length of a flagpole while leaving other causally relevant factors undisturbed, the length of its shadow will change, but changing the shadow's length by changing the elevation of a light source or the angle the pole makes with the ground or in any other way that does not involve directly changing the flagpole's length will not result in a change in the pole's length. In this sense, the length of the shadow is counterfactually dependent on (and is explained by) the length of the pole and not vice versa. Again, changing whether there is a blue spot on the cue ball will change not change the subsequent motion of the balls but changing their linear momentum will. In this sense, the subsequent motion counterfactually depends on (and is explained by) the momentum but not the spot.

This view of the connection between explanation and counterfactual dependence allows us to deal with a puzzle that will have occurred to the alert reader. On the one hand, derivations from laws or other general principles seem to play an explanatory role in many areas of science. On the other hand, (Ex3) and (Ex4) seem to show that not all such derivations are explanatory and (Ex1) seems to show that not all explanations take the form of derivations. We may resolve this puzzle by rethinking the role of derivational structure in explanation. According to the DN model, the role of derivation from a law is to show that the explanandum phenomenon was to be expected. I suggest instead that explanations explain in virtue of conveying information about patterns of counterfactual dependence. Derivation from a law is sometimes a very effective way of conveying such information, as when a derivation of the subsequent motion of the cue balls from the conservation of linear momentum and their prior momenta shows us in a very detailed and fine grained way exactly how the subsequent motion of the balls would have been different in various ways if their prior momentum had been different in various ways. However, not all derivations from laws convey such information about counterfactual dependence and when they do not, as in the case of (Ex3), there is no explanation. Moreover, there are other ways of conveying such

counterfactual information besides explicit derivation and as long as information is conveyed, one has an explanation. Thus, (Ex1) tells us about the counterfactual dependence of the ink tipping on the knee impact and is explanatory for just this reason – we need not see it as explanatory in virtue of instaniating an implicit DN structure, which in any event is not sufficient for explanatoriness in the absence of counterfactual dependence. Other representational devices such as diagrams and graphs similarly convey information about counterfactual dependence without consisting of explicit derivations.

There are many counterfactual theories of causation in the philosophical literature – David Lewis' theory (1973) is probably the best known.[8] For the most part, however, philosophers of science have been unwilling to make extensive use of counterfactual notions in developing theories of explanation. This attitude is partly due to suspicion that counterfactuals fail to meet the empiricist strictures described at the start of this chapter, but it has been exacerbated by features of the very influential semantics for counterfactuals developed by Lewis. Although the semantics is a wonderful achievement, its appeal to trade-offs along different dimensions of "similarity" across "possible worlds" and to "miracles" that violate laws of nature leaves it opaque how counterfactual claims can be tested by ordinary empirical evidence and seems to have little contact with scientific practice. The result has been to make counterfactuals look scientifically disreputable. Recently, however, this situation has changed. Judea Pearl and others – see especially Pearl (2000) – drawing on a substantial preexisting traditions in disciplines like statistics, experimental design, and econometrics have provided rigorous formal frameworks for exploring the connection between causation and counterfactuals. They have also emphasized the very close connection (gestured at above) between counterfactuals and experimentation, and have explored the ways in which even when experimentation is not possible, statistical evidence may be brought to bear on causal claims; in the latter connection, see especially, Spirtes et al. (1993). Although I lack the space to defend this judgment, I think this work goes a long way toward making counterfactuals and accounts of explanation and causation based on counterfactuals scientifically respectable. The task then becomes one of working out in detail how various causal and explanatory notions can be captured within this counterfactual/experimentalist framework – work of this sort is already underway[9] and, in my judgment, represents one of the most promising future directions in the theory of explanation. I will also add the prediction that the best work in this area will make use of formal machinery like systems of equations and directed graphs – machinery that is both richer than representational devices standardly employed by philosophers (logic, probability theory unsupplemented by anything else) and closer to the machinery employed by science itself. Neither logic nor probability theory by themselves can capture the modal and counterfactual elements that are central to explanation.

"Laws of nature" is also a topic on which much work remains to be done. There are many questions that need to be answered. Which if any of the traditional criteria for lawfulness can be reformulated in a defensible way? Is it possible to draw

a relatively sharp distinction between laws and non-laws at all and, if so, does this distinction coincide with the distinction between those generalizations that can figure in explanations and those that cannot, as DN theorists claim? If there is no clear distinction, what follows for the theory of explanation? What are the advantages and disadvantages of thinking of the generalizations of the special sciences as laws even though they lack many of the features traditionally assigned to laws? My suspicion is that progress on these issues will require abandoning the "all As are Bs" framework for representing laws traditionally favored by philosophers in favor of a focus on examples of real laws, which are represented by equations of various sorts which have a much richer structure.

The issue of reductionism also merits rethinking. A great deal of work on explanation, including the accounts described above, seems animated by the assumption that, without a full reduction, no interesting progress has been made. This attitude is not self-evidently correct. Some non-reductionist theories of causation/explanation (e.g., *c* explains *e* if *c* produces *e*, with no further account of "production") do seem completely unilluminating. But not all non-reductive theories are trivial in the way just illustrated. Non-reductive theories can be interesting and controversial in virtue of conflicting with other reductive or non-reductive theories and suggesting different assessments of particular explanations. For example, even if the CM model fails to fully meet empiricist strictures, it will still disagree with counterfactual theories (including non-reductive versions of such theories) in its assessment of explanations that appeal to action at a distance or otherwise fail to trace continuous causal processes, since counterfactual theories presumably will regard such explanations as legitimate. Relatedly, even if we opt for a non-reductive account of some notion within the circle of concepts that includes "cause," "counterfactual," etc., we still face many non-trivial choices about exactly how this notion should be connected up with or used to elucidate other notions of interest – choices that can be made in more or less defensible ways. Finally, even in the absence of a fully reductive account of explanation, it may be possible to show how particular explanatory/causal claims can be tested by making use of other particular causal claims and correlational information. My own view is that, in their enthusiasm for reductive accounts, philosophers have often misdescribed the structure of the explanatory claims they have hoped to reduce. I also think that many of the empiricist constraints imposed on accounts of explanation have been abandoned elsewhere in philosophy and have little justification. Regardless of whether this is correct, the entire subject would benefit from a more explicit discussion of the rationale for the constraints that are standardly imposed.

Notes

1 Woodward (1989) argues it is a misconception that statistical theories explain individual outcomes. Instead, they explain features of probability distributions such as expectation values.

2 For example, the paradigmatically accidental generalization "All the balls in this urn are red" arguably supports the counterfactual "If a ball were drawn from this urn, it would be red." If we want to use support for counterfactuals to distinguish laws, we need to be more precise about which counterfactuals are supported by laws but not by accidental generalizations. Criterion (5) is arguably satisfied by accidental cosmological uniformities such as the generalization that at a sufficiently large scale the mass distribution of the universe is uniform, since these play a unifying role in cosmological investigation. Several of the objections to unificationist theories of explanation discussed below also appear to tell against this criterion.

3 Virtually all recent treatments of confirmation, whether Bayesian or non-Bayesian, agree that "positive instances" by themselves never confirm generalizations, whether lawful or accidental. Instead, it is only in conjunction with background assumptions that positive instances or any other form of evidence can be confirming. Once this is recognized, it becomes clear that in conjunction with the right background assumptions, accidental generalizations are just as confirmable by a limited number of instances as lawful generalizations. For example, in conjunction with the information that an appropriate small sample has been drawn randomly from the US population, the sample can accidental generalizations about political attitudes in that population.

4 Some readers may respond that (L) is not a bona-fide law but this just illustrates again that defense of the DN model requires a more adequate account of laws.

5 See especially Cartwright (1979) and Spirtes et al. (1993).

6 For more on this theme, see Woodward (1989).

7 Kitcher's implausible assumption that there is a single OD pattern of explanation also invites further comment. While the assumption may make little difference to the particular example under discussion, for reasons described in Barnes (1992), it raises the important issue of whether there are non-arbitrary criteria for counting or individuating patterns of argument.

8 My own defense of a counterfactual theory of explanation can be found in Woodward (1984) and Woodward (2000).

9 In addition to Pearl (2000) see, for example, Hitchcock (2001).

References

Barnes, E. (1992): "Explanatory Unification and the Problem of Asymmetry," *Philosophy of Science*, 59, 558–71.

Bromberger, S. (1966): "Why Questions," in R. Colodny (ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, Pittsburgh: University of Pittsburgh Press, 86–111.

Cartwright, N. (1979): "Causal Laws and Effective Strategies," *Nous*, 13, 419–37.

Friedman, M. (1974): "Explanation and Scientific Understanding," *Journal of Philosophy*, 71, 5–19.

Hempel, C. (1965): *Aspects of Scientific Explanation and Other Essays in Philosophy of Science*. New York: Free Press.

Hitchcock, C. (1995): "Discussion: Salmon on Explanatory Relevance," *Philosophy of Science*, 62, 304–20.

Hitchcock, C. (2001): "The Intransitivity of Causation Revealed in Equations and Graphs," *The Journal of Philosophy*, xcviii (6), 273–99.

Kim, J. (1999): "Hempel, Explanation, Metaphysics," *Philosophical Studies*, 94, 1–20.

Kitcher, P. (1989): "Explanatory Unification and the Causal Structure of the World," in W. Salmon and P. Kitcher (eds.), 410–505.

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

Mitchell, S. (1997): "Pragmatic Laws," *PSA 96*, Supplement to *Philosophy of Science* 64(4), S468–S479.

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

Salmon, W. (1971): "Statistical Explanation and Statistical Relevance," in W. Salmon (ed.), *Statistical Explanation and Statistical Relevance*, Pittsburgh: University of Pittsburgh Press, 29–87.

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

Salmon, W. and Kitcher, P. (eds.) (1989): *Minnesota Studies in the Philosophy of Science, Vol 13: Scientific Explanation*. Minneapolis: University of Minnesota Press.

Scriven, M. (1962): "Explanations, Predictions and Laws," in H. Feigl and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science, volume III*, Minneapolis: University of Minnesota Press, 170–230.

Spirtes, P. Glymour, C. and Scheines, R. (1993): *Causation, Prediction and Search*. New York: Springer-Verlag.

Woodward, J. (1984): "A Theory of Singular Causal Explanation," *Erkenntnis*, 21, 231–62.

Woodward, J. (1989): "The Causal Mechanical Model of Explanation," in W. Salmon and P. Kitcher (eds.), 357–83.

Woodward, J. (2000): "Explanation and Invariance in the Special Sciences," *British Journal for the Philosophy of Science*, 51, 197–254.

The Blackwell Guide to the Philosophy of Science

Edited by

Peter Machamer and Michael Silberstein

BLACKWELL Publishers

First published 2002

2 4 6 8 10 9 7 5 3 1

Blackwell Publishers Inc.
350 Main Street
Malden, Massachusetts 02148
USA

Blackwell Publishers Ltd
108 Cowley Road
Oxford OX4 1JF
UK

Library of Congress Cataloging-in-Publication Data has been applied for.

ISBN 0-631-22107-7 (hardback); 0-631-22108-5 (paperback)

British Library Cataloguing in Publication Data
A CIP catalogue record for this book is available from the British Library.

Typeset in 10 on 13 pt Galliard
by Best-set Typesetter Ltd., Hong Kong
Printed in Great Britain by T.J. International, Padstow, Cornwall

This book is printed on acid-free paper.