Explanatory Generalizations, Part II: Plumbing Explanatory Depth*

1. Introduction

Some explanations are deep and powerful: Newton's explanation of the tides, Maxwell's explanation of the propagation of light, Einstein's explanation of the advance of the perihelion of Mercury. Other explanations, while deserving of the name, are superficial and shallow: Bob lashed out at Tom because he was angry, the car accelerated because Mary depressed the gas pedal with her foot, the salt dissolved because it was placed in water. We take this intuition to be very natural and widely shared. Yet in the vast philosophical literature on explanation, there have been precious few attempts to give any systematic account of this notion of explanatory depth. In this paper, we will provide such an account from within the framework of the manipulationist account of explanation presented in a companion paper (Woodward and Hitchcock forthcoming, hereafter referred to as EG1; see also Woodward 1997a, 2000).

We believe that the absence of any adequate theory of explanatory depth is no accident. According to most theories of explanation, explanations appeal (at least tacitly) to generalizations of some sort. For example, in Hempel's Deductive-Nomological (D-N) theory of explanation (Hempel and Oppenheim 1948, Hempel 1965a), explanations must appeal to true, lawlike generalizations — i.e., to laws. A generalization is a proposition that is general in the following sense: it describes more than just the actual properties of the particular system that is the focus of explanation. This suggests a natural approach to the problem of explanatory depth: an explanation is deeper insofar as it makes use of a generalization that is more general. We will ultimately endorse a version of this strategy. We will argue, however, that traditional approaches to explanation have been unable to exploit this strategy because they have focused on the wrong sort of generality: generality with respect to objects or systems other than the one that is the focus of explanation. The right sort of generality is rather generality with respect to other possible properties of the very object or system that is the focus of explanation.

The importance of this sort of generality follows from the claim that explanation has to do with the exhibition of patterns of counterfactual dependence describing how the systems whose behavior we wish to explain would change under various conditions. This claim is articulated and defended in detail in EG1; we present a brief overview in section 2. In section 3, we provide our account of explanatory depth in terms of the range of invariance of a generalization. In particular, we describe a number of different ways in which one explanation may be deeper or more powerful than another. In the remaining sections, we explore the differences between our account and two other influential treatments of explanation:[1] Kitcher's unificationist account (Kitcher 1981, 1989) and Lewis's causal account (1986a). We argue in particular that neither theory provides an adequate account of the notion of explanatory depth.

2. Explanatory Generalizations

An explanation has the following canonical form:

(1) $X_1 = x_1, \ldots, X_n = x_n$
$Y = g(X_1, \ldots, X_n)$
$\therefore Y = y = g(x_1, \ldots, x_n).$

The first two lines comprise the explanans, the final line the explanandum. $X_1, \ldots X_n$ are the explanans variables, and Y the explanandum variable. A variable is a determinable property of an object or system: its mass, its electric charge, whether it is a raven, etc. The second line expresses a relationship between the explanans and explanandum variables; it is what we call the explanatory generalization of explanation (1). We impose two further constraints upon an explanation. First, the assignments of values to the explanans and explanandum variables must be true (or approximately true) of the object or system in question. Second, the explanatory generalization

must be invariant under testing interventions. In order for a generalization to be explanatory in our sense, it need not be a law, and indeed it need not be an exceptionless regularity. There may be other systems of the type in question for which $Y \neq g(X_1,\ldots, X_n)$; there may even be systems such that $X_1 = x_1,\ldots, X_n = x_n$ and $Y \neq y$.

In EG1, we provide a detailed account of what it means for a generalization to be invariant under testing interventions. The basic idea is that there must be some true counterfactuals of the following form: if it had been the case that $X_1 = x_1',\ldots, X_n = x_n'$, then it would have been the case that $Y = y' = g(x_1',\ldots, x_n') \neq y$. (If $y = y'$, then we say that the relevant intervention was not a testing intervention.) These counterfactuals are non-backtracking counterfactuals in the sense of Lewis (1979). We do not, however, endorse Lewis's account of such counterfactuals in terms of a metric of similarity over possible words that is characterized in terms of big and small 'miracles'. Rather, the counterfactuals are to be thought of as describing hypothetical situations in which the antecedents are made true by idealized interventions or manipulations of the explanans variables. It is precisely these sorts of idealized interventions that we aim to approximate when we conduct controlled experiments. The key feature of such interventions is that they do not exert any causal influence on the explanandum variable Y except through their effect, if any,[2] on the explanans variables.

Explanatory generalizations allow us to answer what-if-things-had-been-different questions: they show us what the value of the explanandum variable depends upon. For example, suppose that the height (Y) of a particular plant depends upon the amount of water (X_1) and fertilizer (X_2) it receives according to the following formula:

(2) $\quad Y = a_1X_1 + a_2 X_2 + U$

where U reflects unknown sources of error. (2) will fall far short of the normal standards for being a law of nature. Nonetheless, suppose that for some ΔX_1 and ΔX_2 (2) correctly 'predicts' that if X_1 and X_2 had been changed by those amounts, then the height of the plant would have changed by (approximately) the amount $a_1\Delta X_1 + a_2\Delta X_2$. Then (2) would qualify as explanatory according to our account, since it gives us information about how the height of the plant depends upon the amount of water and fertilizer that it receives.

3. Degrees of Invariance and Explanatory Depth

We argued in EG1 that the traditional distinction between laws and accidental generalizations does not do justice to the role played by generalizations in explanation. Moreover, the traditional distinction involves an exhaustive dichotomy of true generalizations — a true generalization is either a law, in which case it is explanatory, or it is accidental, in which case it is not explanatory. There are no other possibilities. However, explanatory generalizations may differ in degree of invariance. For example, the range of conditions over which the regression equation (2) is invariant is intuitively rather narrow in comparison with fundamental physical laws. An initial step in the right direction thus would be to abandon the law/accident dichotomy, and replace it with an alternative framework that involves a threshold above which there is a continuum that admits of degrees. Among those generalizations that are invariant, some will be more invariant than others, and they will correspondingly provide deeper explanations. For example, the low-level generalization (2) relating water and fertilizer to plant height strikes us as explanatory, but only minimally so: the explanations in which it participates are shallow and relatively unilluminating. If we had a theory — call it (T) — describing the physiological mechanisms governing plant growth it would provide deeper explanations. Such a theory would presumably be invariant under a wider range of changes and interventions than (2); that is, we would expect (T) to continue to hold in circumstances in which the relationship between height, fertilizer and water departed from the linear relationship (2). In just the same way, the van der Waals equation will be invariant under a

wider range of interventions than the ideal gas law and hence will be more explanatory, General Relativity will be more explanatory than Newtonian gravitational theory, and so on.

Note that this explanatory spectrum will not necessarily be one that extends from accidental generalizations at one extreme, to laws at the other. As we argued in EG1, some generalizations that have traditionally been considered laws are not genuinely explanatory. Moreover, the traditional distinction between laws and accidental generalizations applies only to exceptionless generalizations. Treating these two categories as endpoints of an explanatory continuum thus falsely suggests that only exceptionless generalizations make it into the continuum at all. As we argued in EG1, exceptionlessness is not necessary for a generalization to be explanatory.[3]

We have been speaking thus far in an informal way of one generalization's being 'more invariant' and hence figuring in deeper explanations than another. We turn now to a more systematic exploration of what this might involve. Let G be a generalization that includes X as one of its explanans variables, and suppose that G is invariant under interventions on the value of X within the range R. Suppose that G' is a different generalization, that purports to explain the same outcome. Then the following are ways in which G' might more invariant than G.

1. G' also includes X as an explanans variable, and is invariant under interventions on X within range R, but G' yields more accurate values for the explanandum variable within that range (even though, ex hypothesi, G was 'approximately true' within this range). This type of improvement is often achieved, for example, by an increase in the accuracy of measurement of a physical constant. Consider the following rendering of Galileo's law of free fall (G): $h = (4.9m/s^2)t^2$, where h is the height (in meters) from which the object in question falls, and t is the time (in seconds) that it takes to fall. A generalization (G') that uses a more accurate value for the acceleration field in some particular location will improve upon G in just the way described.

2. As before, G' includes X, but it is invariant under interventions on X within range R', which strictly contains R. In this case, G' is invariant under interventions on X that G is not invariant under.

It might be natural to regard the first type of improvement as an improvement in accuracy, which is an explanatory virtue distinct from depth; while regarding the second type of case as an increase in invariance per se. Note, however, that these two cases frequently occur together. For example, Newton's laws are highly accurate for low velocities; that is, Newton's laws, when applied to some object with a velocity that is very small compared to that of light, will be invariant under a range R of interventions on that velocity, so long as R includes only relatively small values of velocity. The special relativistic correction to these laws has two interrelated effects: the new generalizations are more accurate within R (even though Newton's laws were 'approximately true' here); and they are invariant under a wider range of interventions on the velocity of the object in question.

3. The case is somewhat more complicated if G and G' have ranges of invariance for X that are partially or totally disjoint. If the actual values of the variables fall within the range of invariance of both G or G', it may be reasonable to prefer G' if it is more accurate within the region of overlap, or if the actual values of the variables fall more squarely within the range of invariance for G'. For example, if G is invariant under interventions that lower the value of X, but not under interventions (or under very few interventions) that raise the value of X, whereas G' is invariant under interventions that both raise or lower the value of X, then it may be reasonable to prefer G' to G.

4. G will be explanatorily deficient if its range of invariance R is too disjoint. Consider for example, Galileo's law of free fall (G') that relates the amount of time it takes an object to fall to earth to the height from which it was dropped: $h = at^2$. Now consider the following 'Goodmanized' version of this law (G): $h + (h - h_0)(h - h_1)...(h - h_n) = at^2$. Suppose that the object whose time of fall we wish to explain was in fact dropped from a height of h_0. Then this new relationship will be invariant under testing interventions that change the value of h to $h_1, \ldots, h_n$. In this case, we are inclined to conclude not merely that G is less explanatory that G', but that G is not explanatory at all. The problem lies in the disjoint nature of the set of values of h for

which (G) holds. In order to get to a testing intervention under which (G) is invariant, we must 'skip over' a set of testing interventions under which (G) is not invariant. Any sufficiently wildly oscillating function of X is bound to hit the right values of the explanandum variable at a number of points; that should hardly lead us to accept such a function as explanatory. Such a generalization is rather like the broken watch that has the dubious virtue of telling exactly the right time twice per day.

A natural way of handling this sort of case would be to require that an explanatory relationship be invariant under all testing interventions that change the values of variables to new values within some neighborhood of the actual values. We might say that such a relationship is stable for the actual values of the variables that figure in it. (G) fails because any neighborhood of $h = h_0$ will contain values of h such that (G) is not preserved when h is set to those values by interventions.

5. It may the case that whether G continues to hold under changes that set X equal to x depends not merely upon whether x is in R, but upon how X is set to x. There are two different types of case — or perhaps two different ways of thinking about the examples — that fall into this category. In the case of some generalizations, some ways of changing the value of X may not count as interventions at all, roughly because causal features of the process by which the value is brought about exert an independent effect on the explanandum variable in G. This would be the case, for example, if the water received by a plant was delivered by a high pressure hose in a way that damaged the plant and disrupted (2). Cases of this sort are not cases in which (2) fails to be invariant under an intervention, since this sort of delivery of water fails to qualify as an intervention; but they nonetheless indicate that the range of interventions under which (2) is invariant is restricted in a certain way.

In other cases, we may be willing to regard various manipulations that fix the value of X within the range R as interventions, but the relationship R is invariant for only some such interventions. For example, it is clear that whether (2) holds will be sensitive to the way in which water is delivered over time: a hypothetical manipulation that dumps large quantities of water on the plant on the last day of the growing season will not cause the plant to become dramatically taller, and will have a quite different effect on height than a manipulation that consists in providing the plant with the same total amount of water but in a way that is distributed more evenly over the growing season. Here we may wish to regard both manipulations as interventions and say that (2) is invariant under the second sort of intervention but not the first. Intuitively, the variables figuring in the generalization are genuine causes of the plant's height, but they are too 'coarse-grained' for (2) to be invariant under a wide range of interventions on those variables.

The boundary between these two sorts of cases is quite fuzzy; typically, however, little will turn on whether specific examples are assimilated to one or to the other. Both of the above examples indicate restrictions on the range of interventions under which (2) is invariant, and it matters little whether we think of these restrictions as arising because interventions exist under which (2) fails to be invariant or whether we instead think of various changes which would disrupt (2) as failing to qualify as interventions. In either case, it is clear that we could improve upon the original generalization (2) by replacing it with a generalization that is less sensitive to the way in which the values of the variables figuring in it are changed; such a generalization will be invariant under a wider range of interventions than (2). For example, we might replace the variable X_1 (total amount of water) with a series of variables representing the amount of water received by the plant during each week of the growing season. This revised generalization would reflect, for instance, the differential effects of water earlier in the growing season rather than later, the effects of long periods without water, and so on. Ideally, one would like to formulate generalizations that are not sensitive at all to the ways in which the values of the variables figuring in them are produced.

6. Even though G may be invariant under some range of interventions on the variables figuring in that relationship, it may be highly sensitive to background conditions, such that it would fail to hold if they were changed in a number of ways. If G' is less sensitive to background conditions, then we might be inclined to view G' as more explanatory than G.

While intuitively plausible, this suggestion raises a puzzle. In EG1 we distinguished between invariance under changes in the value of a variable figuring in a generalization, and invariance under changes in background conditions, and concluded that the latter sort of invariance did not suffice for even minimal explanatory power. The problem was that any generalization (that is true or approximately so of the system at hand) will be invariant under a great many changes in background conditions (such as changes in the price of tea in China). So there appears to be a tension (although not an outright contradiction) between two claims: invariance under changes in background conditions does not render a generalization explanatory; yet greater invariance under changes in background conditions can render one generalization more explanatory than another. We resolve this tension by claiming that all interesting cases of this type can be assimilated to case 7 below. Briefly, if G is sensitive to changes in background conditions, that is because it has left out some variable(s) upon which the explanandum variable depends; whereas if G' holds under a variety of background conditions, that will be because G' has already incorporated many of the relevant variables into the generalization.

7. G' makes explicit the dependence of the explanandum on variables treated as background conditions by G. For example, Galileo's law of free fall (G) expresses a relationship between the height from which an object is dropped and the time it takes to fall. This generalization is invariant under some interventions on the height from which the object is dropped, but it would fail to hold if the object were dropped from a height that is large in relation to the earth's radius or if it were dropped from the surface of a massive body of proportions different from those of earth (such as Mars). Newton's second law together with his law of gravitation entail a generalization (G') that also allows us to compute the time it would take an object to fall a certain distance. (G'), however, is not restricted in the way that Galileo's law is: it will remain invariant under changes in the mass and radius of the massive body upon which the object is dropped. It achieves this greater range of invariance by explicitly incorporating the mass and radius of the planet (or whatever) into the generalization as variables.[4]

As a second example, consider the Hardy-Weinberg law of population genetics, which shows how the population frequencies of genotypes in one generation depend upon the frequencies of the individual genes (alleles) in the previous generation. This 'law' only holds under a very restrictive set of assumptions: there must be no migration into or out of the population and no mutation, there must be random mating, random assortment of genes during meiosis, and no difference in reproductive fitness conferred by the genes. Population genetics allows for the construction of more complex equations that show how the population frequency of genotypes can depend upon factors other than initial gene frequencies, factors such as migration, mutation, meiotic drive, fitness, and so on.

In these sorts of cases, claims about the invariance of a relationship under changes in background conditions are transformed into claims about invariance under interventions on variables figuring in the relationship through the device of explicitly incorporating additional variables into the relationship. For example, an intervention that increases the mass of the earth would count as an intervention on background conditions with respect to Galileo's law, but as an intervention on a variable explicitly figuring in Newton's laws. This is, perhaps, the most fundamental way in which one generalization can provide a deeper explanation than another. At the heart of explanation is showing what the explanandum depends on. If an explanandum depends on some variable, a generalization that explicitly describes this dependence achieves this aim more fully than a generalization that does not make this dependence explicit.

Note that the various ways in which one generalization may be deeper than another may sometimes compete — explanatory depth is not one-dimensional. Consider, for example, the relationship (2) between the height of a plant and the amount of water and fertilizer it is given. There may well be deeper theories of plant growth that exhibit the dependence of the plant's height on a great many other variables besides. It may be, however, that these deeper theories do not exactly reproduce (2) when the appropriate boundary conditions are specified. In other words, there may be a tension between the desiderata for accuracy (case 1) and explanatory depth (case 7). It is often the case in the biological and social sciences that low-level generalizations that are formulated primarily on the basis of straightforward inductive evidence are more accurate within

their respective domains than are the deeper explanatorily generalizations of those fields. We see this, for example, in the contrast between old-fashioned institutional economics which is often descriptively accurate but explanatorily very shallow, and modern industrial organization theory, which is much more illuminating from the point of view of explanation but abstracts from details of particular firms and markets in a way that sacrifices descriptive accuracy. Cartwright (1983) contains a sustained defense of this position for the generalizations of physics. Cartwright distinguishes between phenomenological laws — such as laws describing the exponential decay rates of unstable isotopes — which are formulated primarily on the basis of direct experimental evidence, and fundamental laws such as Schrödinger's equation which are thought to have deep explanatory power. She argues that phenomenological laws are never merely the result of applying boundary conditions to fundamental laws: so-called derivations of phenomenological laws from fundamental laws always involve empirical idealizations, such as ignoring all but a finite number of factors that contribute to a system's Hamiltonian, and mathematical approximations such as discarding small terms in infinite sequences and substituting tractable equations for intractable ones. Thus the result of applying a fundamental law to some particular system is often less descriptively accurate than an empirically derived phenomenological law. Whether or not this is a correct account of generalizations in physics, Cartwright is correct in thinking that the demands for accuracy and for explanatory depth can sometimes pull in quite different directions.

Our account of the variety of ways in which one generalization may be more invariant than another allows us to understand better why the traditional nomothetic approach to explanation has been unable to provide an adequate account of explanatory depth. As we argued in EG1, the nomothetic approach has focused on a particular kind of generality: generality with respect to objects or systems other than the one whose properties are being explained. This type of generality is ill-suited to the project of developing a systematic account of explanatory depth for a number of reasons.

First, there is a technical difficulty. Laws, as traditionally understood, are universal generalizations. This means that a law asserts that a particular material conditional holds of everything. In this regard, it is hard to see how one explanatory generalization could be more general than another. A natural response to this problem is to suggest that every law has a certain 'scope', a set of objects or systems that fall under the antecedent of the law. We could then say that one law is more general than another if it has a wider scope.

The difficulty with this suggestion is that scope has little to do with the aims of explanation as articulated by Hempel. Hempel writes:

> …a D-N explanation answers the question 'Why did the explanandum-phenomenon occur?' by showing that the phenomenon resulted from certain particular circumstances, specified $C_1, C_2,\ldots C_k$, 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. (1965a, 337)

An explanation that cites a generalization having narrow scope serves this purpose just as well as one that cites a generalization of wider scope (so long as the system in questions falls within the scope of each generalization) By contrast, on our account the aim of explanation is to provide the resources for answering what-if-things-had-been-different questions by making explicit what the value of the explanandum variable depends upon. We have shown in detail how generalizations that are invariant under a wider range on interventions better serve this aim. D-N explanations appealing to generalizations with narrow and wide scope will do equally good jobs of showing that their explananda 'were to be expected'; by contrast, explanations can differ in the extent to which they can be used to answer a range of what-if-things-had-been-different questions, and such differences connect to differences in explanatory depth.

Finally, increased scope does not always correspond to explanations that are intuitively deeper. The conjunction of Galileo's law and the Boyle-Charles law has wider scope than either conjunct

taken separately: it makes interesting predictions about both falling objects and gases, while each conjunct taken separately makes predictions about only one of these classes of objects. Yet if we want to explain why an object took three seconds to fall, an explanation that cites the gerrymandered conjunctive law is intuitively no deeper than the one that cites Galileo's law alone. Our account validates this intuition: the conjunctive law does not increase our understanding of the variables upon which the falling time depends. This judgment is in turn reflected in the distinction between 'other object' and 'intervention' counterfactuals, described in EG1. The conjunction of Galileo's law and the Boyle-Charles law does support counterfactuals that are not supported by Galileo's law alone: for instance, counterfactuals about the behavior of ideal gasses. Nonetheless, the conjunctive law provides no more information about what will happen under interventions on variables affecting the falling time of objects in free fall than does Galileo's law alone.

A closely related observation is that our account, and in particular point 7 above, can be used to show how generalizations may sometimes be used to explain other generalizations. This was a notorious difficulty for the D-N model of explanation. The problem, as noted by Hempel and Oppenheim in their famous footnote 33 (1948 [Hempel 1965b], p. 273), is how to distinguish genuine explanations of generalizations, such as a derivation of (an approximation of) Galileo's law of free fall from Newton's laws, from spurious explanations, such as a derivation of Galileo's law from the conjunction of Galileo's law and the Boyle-Charles law. Our account neatly draws the distinction on the grounds that Newton's laws are invariant under testing interventions with respect to Galileo's law, whereas the conjunctive Galileo-Boyle-Charles law is not. That is, interventions on the values of variables figuring in Newton's laws, such as the mass and radius of the earth, would result in various alternatives to Galileo's law. Newton's laws show how the truth of Galileo's law depends upon the values of certain variables. By contrast, there is no intervention on the value of a variable figuring in the conjunctive Galileo-Boyle-Charles law that would lead to Galileo's law being false. The conjunctive law in no way shows what the truth of Galileo's law depends on. Unlike the conjunctive law, Newton's laws provide more information about what would happen under interventions on variables affecting time of fall than does Galileo's law alone. Thus while there is a sense in which both Newton's laws and the conjunction of Galileo's and the Boyle-Charles law are 'more general' than Galileo's law alone (both have greater scope in the sense of enabling predictions about more systems), this sort of generality does not in itself provide for deeper explanations. The kind of increase in generality that matters for increased explanatory depth is the very specific kind exhibited in the relationship between Newton's laws and Galileo's, in which we are shown what the truth of Galileo's law depends on.

We should note, however, that by no means everything that we may wish to count as an explanation of a generalization fully fits the pattern we have been describing. For example, it is often argued that the stability of planetary orbits depends (mathematically) upon the dimensionality of the space-time in which they are situated. This accords reasonably well with our idea that explanations provide answers to what-if-things-had-been-different questions: the derivation may tell us what would happen if space-time were five-dimensional and so on. Mark Steiner has argued that genuinely explanatory mathematical proofs have this character:

> My proposal is that an explanatory proof makes reference to a characterizing property of an entity or structure mentioned in the theorem, such that from the proof it is evident that the result depends on the property. It must be evident, that is, that if we substitute in the proof a different object of the same domain, the theorem collapses; more, we should be able to see as we vary the object how the theorem changes in response. (Steiner 1978, 143.)

However, it seems implausible to interpret such derivations as telling us what will happen under interventions on the dimensionality of space-time, etc. One natural way of extending our position would be as follows: all explanations must answer what-if-things-had-been-different questions. When a theory tells us how Y would change under interventions on X, we have (or have material for constructing) a causal explanation. When a theory or derivation answers a what-if-things-had-been-different question, but we cannot interpret this as an answer to a question about what would happen under an intervention, we may have a non-causal explanation of some sort. This accords

with intuition: It seems clear that the dependence of orbital stability upon dimensionality or the dependence of a theorem on the assumptions from which it is derived is not any sort of causal dependence. In this paper, our focus is on causal explanation.[5]

4. Kitcher on Explanatory Unification

In this section and the next, we will examine two theories of explanation that bear some affinities with our own: Philip Kitcher's unificationist account and David Lewis's causal account. We will argue that neither of these accounts can supply an adequate account of explanatory depth.

According to Kitcher, the fundamental idea of the unificationist approach is that

> Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and, in demonstrating this, it teaches us how to reduce the number of types of facts we have to accept as ultimate (or brute). (1989, 432.)

Put slightly differently the idea is that successful explanations unify by allowing us to deduce a range of different facts from some much smaller number of fundamental assumptions by repeatedly using the same patterns of derivation.

The unificationist approach bears some similarities to our approach; in particular, both approaches take it as an essential feature of explanatory generalizations that they apply to a number of different cases. But the two accounts differ fundamentally in what they take the relevant cases to be. In particular, the unificationist approach understands the relevant cases in terms of the notion of scope while we focus instead on range of invariance. A generalization can have very wide scope while being invariant only under a narrow range of interventions or indeed without being invariant under any interventions at all. Conversely, a generalization can have narrow scope while being invariant under a wide range of interventions.

By way of illustration, consider two brief examples. The generalization (K) that specifies that for each spatio-temporal region of the universe, the microwave radiation background left over from the big bang is 3°K has very wide scope. (K) is a unifying generalization that could be used , over and over again, in the derivation of many different phenomena. Nonetheless, it does not follow from this fact that (K) is invariant over a wide range of interventions. In fact it is not clear that there are any well-defined testing interventions with respect to (K). In addition, contemporary cosmological theorizing suggests that whether or not (K) holds is extremely sensitive to the precise initial conditions that obtained in the very early universe. If these were even slightly different (K) would not hold. (K) describes an extremely pervasive uniformity but pervasiveness has to do with scope, not invariance. On our account (K) is not an explanatory generalization. Unificationist accounts seem to reach the opposite conclusion.

As a second illustration, imagine two different neural circuits N_1 and N_2. N_1 is, as biologists say, highly conserved — it is found in many different kinds of organisms, as diverse as snails and human beings, and the same generalizations describe its behavior in each case. By contrast N_2 is found only in a certain species of snail. The generalizations governing the behavior of N_1 have much greater scope than the generalizations governing N_2, but again it does not follow that they are invariant under a wider range of interventions. It is entirely possible that the generalizations governing N_2 and those governing N_1 are invariant under exactly the same interventions on neural structure. While the unificationist account seems to yield the conclusion that the generalizations governing N_1 provide more unified and hence better or deeper explanations than the generalizations governing N_2 simply in virtue of applying to more organisms (or more different kinds of organisms) our account avoids this unintuitive conclusion.

There is another fundamental problem with the unificationist account. We have been assuming so far that on this account explanations can differ in degree of unification and that the more unified an explanation is the greater its explanatory depth. An explanation can provide less unification than some alternative, and hence be shallower, but still qualify as somewhat explanatory. This is,

we submit, the natural way of connecting unification and explanatory depth on a unificationist account. Unfortunately, Kitcher's treatment of a number of familiar puzzle cases requires rejection of this assumption. And without it, it is hard to see how to provide any plausible treatment of explanatory depth within a unificationist framework.

To see the difficulty, recall Kitcher's treatment of the problems of explanatory irrelevance and explanatory asymmetry. Why is it that we cannot appeal to the fact that Jones, a male, has taken birth control pills to explain his failure to get pregnant? According to Kitcher, any explanatory store of which this 'explanation' is a part will be 'less unified' than a competing explanatory store according to which the failure of males to become pregnant is always explained in terms of their gender rather than their ingestion of birth control pills. Similarly, the reason why we cannot explain the height of a flagpole in terms of the length of the shadow it casts is that explanations of lengths of objects in terms of facts about shadows do not belong to the 'set of explanations' which 'collectively provides the best systemization of our beliefs' (1989, p. 430). Quite apart from any other doubts one may have about these claims, they clearly require the idea that explanations that are less satisfactory from the point of view of unification than some alternative are unexplanatory, rather than merely less explanatory than the alternative. However, if we accept this idea, we lose the 'natural' connection between unification and explanatory depth described above. If we have two competing explanations of the same subject matter, one of which is more unified than the other, the second cannot be shallower than the former but nonetheless explanatory; instead it must be completely unexplanatory.

Kitcher's version of the unificationist account thus faces a dilemma. On the one hand, it seems very natural and desirable to say that generalizations and theories can sometimes be explanatory even though alternative deeper (or 'more unifying') explanations are known or discoverable. For example, the regression equation (2) relating quantities of water and fertilizer to plant height is explanatory even though there are far deeper, more biologically grounded explanations of plant growth. Similarly, Galileo's law can be used to explain facts about the behavior of falling bodies even thought it is less explanatorily deep than the laws of Newtonian mechanics, the latter are explanatory even though they are less deep than those of special and general relativity and so on. If we reject this idea, we seem led to the conclusion that only the deepest, most unified theories are explanatory at all; everything else is non-explanatory. This is to completely give up on the idea that there are degrees of explanatory depth .

Suppose, on the other hand, that we agree that if theory T_2 is deeper or more unified than T_1, it does not automatically follow that T_1 is unexplanatory. Then Kitcher's solution to the problems explanatory irrelevance and asymmetry are no longer available: it isn't clear why we shouldn't conclude that an 'explanation' of Jones failure to get pregnant in terms of his ingestion of birth control pills is genuinely explanatory, although shallower than the alternative explanation that invokes his gender .

Intuitively, the problem is that we need a theory of explanation that allows us to capture several different possibilities. On the one hand, there are generalizations and associated putative explanations (like the generalization (3) relating barometric pressure to rain or the generalization in (7) which relates ingestion of birth control pills among men to failure to get pregnant) that are not explanatory at all; they fall below the threshold of explanatoriness. On the other hand, above this threshold there is something like a (multi-dimensional) continuum: a generalization can be explanatory but less deep than some alternative. What we have just seen is that the unificationist account has difficulty capturing simultaneously both of these possibilities — either there is no threshold or there is no continuum. By contrast, the account we have proposed does this in a very natural way. Some generalizations are not invariant under any testing interventions at all and hence are non-explanatory. Other generalizations such as (2) are invariant under some interventions (and answer some what-if-things-had-been-different questions) and hence are above the threshold of explanatoriness. Nonetheless they are less invariant (in the sense captured by (1) - (7) of section 3) than other generalizations and hence less explanatory.

5. Lewis on Causal Explanation

David Lewis has propounded a theory of explanation according to which 'to explain an event is to provide some information about its causal history' (Lewis 1986a, 217).[6] This thesis naturally suggests an approach to the problem of explanatory depth: one explanation is deeper or more powerful than another if it provides more information about causal history. In contrasting his view with the D-N approach, Lewis writes: 'It's not that explanations are things that we may or may not have one of; rather, explanation is something we may have more or less of.' (1986a, 238). This sounds very much in accord with our own claim: generalizations do not divide neatly into laws, which are explanatory, and accidental generalizations, which are not; rather, generalizations can have degrees of invariance and hence degrees of explanatory power. In fact, however, Lewis stresses increases in explanatory information along a different dimension than we do.

As an intuitive aid, imagine some event to be explained as a point at the top of a page, and the various causal chains leading up to that event as lines leading up to that point. Lewis complains that a 'D-N argument presents only one small part — a [horizontal] cross section, so to speak — of the causal history' (1986a, 237). Suppose for example, we want to explain why a particular plant grew to height y just before it was harvested. Suppose, moreover, that generalization (2) is somehow transformed into a genuine law (2') that relates the height of the plant to quantity of water and fertilizer, as well as to other initial conditions. Then we could provide a D-N argument for the height of the plant from (2') and the appropriate initial conditions. Still, this would not be a complete explanation. It is possible to 'interpolate' and 'extrapolate' more causal information. For instance, we could provide information about why the plant received quantity x_1 of water during the growing season, perhaps in terms of meteorological conditions. Or we could explain how the water, fertilizer, etc. affected certain growth processes in the plant, and how these in turn affected the plant's height. That is, we could provide more information along the 'vertical' dimension by describing causes and intermediate effects of the plant's receiving x_1 amount of water.

In section 3 above, we focused primarily on degrees of explanatory power along the 'horizontal dimension'. But we have no objection to Lewis's concern with providing more explanatory information along the vertical dimension. If we have an explanation of the height of a particular plant in terms of the amount of water and fertilizer it received, then we can improve upon it in the various ways suggested in section 3 above. We can also improve upon it by extrapolating or interpolating: explaining (by means of invariant generalizations) why the plant received quantity x_1 of water or why it underwent certain growth processes. Causal explanations of the sort we have described are the building blocks of complete explanations: they can be 'stacked' or 'subdivided' to provide richer causal histories.

Nonetheless, we believe that there is a horizontal component to explanatory depth that is omitted from Lewis's account. That is, even when we restrict ourselves to one of the basic building blocks, we can still distinguish between degrees of explanatory depth in a way that Lewis cannot. In order to appreciate this point, we must turn to Lewis's theory of causation. According to Lewis (1973), the (occurrent) event e is counterfactually dependent upon the (distinct, occurrent) event c if the following counterfactual is true: if c had not occurred, e would not have occurred. In order to avoid certain problems involving pre-emption, Lewis defines causation as the ancestral of counterfactual dependence, and not as counterfactual dependence itself. No matter: relations of counterfactual dependence are still the basic building blocks of causal histories. To provide a (basic) explanation, then, is to provide information about the causal history of the explanandum event, which is to provide information about events upon whose occurrence the explanandum event counterfactually depends.

Lewis's account of explanation thus agrees with ours that explanations provide the resources for answering what-if-things-had-been-different questions. To provide the information that c is part of the causal history of e is (to a first approximation) to provide the information that if c had not occurred, e would not have occurred. However, such information does not enable us to answer very many what-if-things-had-been-different questions: instead each causal claim gives the answer to only one such question. To illustrate this, suppose that the electric field intensity at a particular point is explained in terms of the charge density along a particular wire, using Coulomb's law (this is example (1) of EG1). On Lewis's account, the wire's having the charge

density it does counts as a cause of the field strength, and hence the charge density is explanatorily relevant to the field strength. To state that the charge density was a cause of the field intensity does provide us with the answer to one what-if-things-had-been-different question: it tells us that if that particular charge density had not occurred, that particular field intensity would not have occurred. Unlike the explanation citing Coulomb's law, however, this causal explanation would not tell us anything about how the field intensity would have changed if the charge density had been different in various ways. An explanation that does exhibit this detailed pattern of dependence is for this reason deeper than one that does not. Lewis, by focusing only on counterfactuals involving the occurrence or non-occurrence of specific events, does not have the resources to capture this notion of explanatory depth.

In a recent paper, Lewis has offered a new theory of causation. According to this theory, causation is (the ancestral) of influence, where

> c influences e iff there is a substantial range c_1, c_2,…of different not-too-distant alterations of c (including the actual alteration of c) and there is a range e_1, e_2,…of alterations of e, at least some of which differ, such that if c_1 had occurred, e_1 would have occurred, and if c_2 had occurred, e_2 would have occurred, and so on. (Lewis 2000, 190)

The intuitive idea is that in order for c to count as a cause of e, it needn't be the case that the occurrence of e depends counterfactually upon the occurrence of c: it is enough if the time and manner of e's occurrence depends upon the time and manner of c's occurrence. Consider our example, in which the value of the electric field at a point depends upon the charge density and geometry of a conductor according to Coulomb's law. Translating into Lewis's terminology, c might be the presence of a conductor with a particular charge density and geometry, and e the presence of an electric potential field with a certain intensity and direction. The 'alterations' c_1, c_2,… would be alternative combinations of charge density and geometry while e_1, e_2,… would be alternative values of the electric field potential. Lewis's definition then comes very close to our notion of invariance. The requirement that there must be true counterfactuals of the form ' if c_i had occurred, e_i would have occurred' is analogous to our requirement that an explanatory relationship between two variables be invariant under interventions, and not merely report a correlation. Lewis's requirement that this relation hold for 'not-too-distant' alterations is analogous to our requirement (elucidated in section 3 above) that an explanatory relationship be invariant under some range of interventions in a neighborhood that includes the actual values of the explanans variables. His requirement that the e_i's not all be identical is analogous to our requirement that an explanatory relationship be invariant under testing interventions.

We can only speculate about how or whether Lewis might revise his theory of causal explanation in light of his new theory of causation. One possible revision would be to maintain that explanations provide information about the causal history of the event to be explained, where this information includes information about the patterns of counterfactual dependence that establish that one event 'influences' another. If so, Lewis's theory would closely resemble ours. There would be some differences of detail — in particular, differences about how the relevant counterfactuals are to be understood (see EG1 for details). Nonetheless, this theory would be close enough to ours that Lewis could make use of much of section 3 to explicate the various ways in which one explanation can be deeper and more powerful than another by exhibiting a wider pattern of counterfactual dependence.

There is, however, some reason to think that Lewis might resist this move. In contrasting his causal theory of explanation with the covering law approach he writes:

> we can ask whether information about covering laws is itself part of explanatory information. The covering law theorist says yes; I say no. (Lewis 1986a, 239.)

The idea seems to be that while a covering law may play a role in establishing that an event c is a cause of e, the law does not itself figure in the explanation of e. We conjecture that Lewis might take the same attitude to our invariant generalizations. An invariant generalization expresses a pattern of counterfactual dependence, which can establish that some event c is part of the causal history of e. In this way, the invariant generalization provides indirect information about the causal history of e. Beyond this, however, the generalization would play no further role. It should be clear that we think this would be a mistake. To say that the charge density's being equal to λ is a cause of the field intensity's being E would imply only that some change in the charge density would yield some change in the field intensity.[7] It seems clear to us that a deeper explanation is to be had by specifying just how the field intensity depends upon the charge density. This is what laws and other sorts of explanatory generalizations do. If Lewis were to treat invariant generalizations as he treats covering laws, he would be throwing away just the resources he needs to provide an account of explanatory depth.

8. Conclusion

In this essay, we have argued that the account of explanatory generalizations articulated in a companion paper (Woodward and Hitchcock forthcoming) provides a natural account of explanatory depth. One generalization can provide a deeper explanation than another if it provides the resources for answering a greater range of what-if-things-had-been-different questions, or equivalently, if it is invariant under a wider range of interventions. That is, generalizations provide deeper explanations when they are more general. It is important, however, to understand generality in the right way: generality with respect to hypothetical changes in the system at hand. By focussing on the wrong sort of generality — generality with respect to systems other than the one whose features are to be explained — rival accounts of explanation such as Hempel's D-N model and Kitcher's unificationist theory have been unable to provide adequate accounts of explanatory depth.

REFERENCES

Cartwright, N. (1983) How the Laws of Physics Lie, Oxford: Clarendon Press.

Hempel, C. (1965a) "Aspects of Scientific Explanation," in Hempel (1965b), pp. 331 - 496.

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

Hempel, C., and P. Oppenheim (1948) "Studies in the Logic of Explanation," Philosophy of Science 15: 135 - 175. Reprinted in Hempel 1965b, pp. 245-290.

Hitchcock, C. (1995) "Discussion: Salmon on Explanatory Relevance," Philosophy of Science 62: 304 - 320.

Kitcher, P. (1981) "Explanatory Unification," Philosophy of Science 48: 507 - 531.

Kitcher, P. (1989) "Explanatory Unification and the Causal Structure of the World," In Scientific Explanation, ed. P. Kitcher and W. Salmon, Minneapolis: University of Minnesota Press, pp. 410 - 505.

Lewis, D. (1973) "Causation," Journal of Philosophy 70: 556 - 567. Reprinted with Postscripts in Lewis 1986b, pp. 159 - 213.

Lewis, D. (1979) "Counterfactual Dependence and Time's Arrow," Noûs 13: 455 - 476. Reprinted with Postscripts in Lewis 1986b, pp. 32 - 66.

Lewis, D. (1986a) "Causal Explanation," in Lewis 1986b, pp. 214 - 240.

Lewis, D. (1986b) Philosophical Papers, Volume II, Oxford: Oxford University Press.

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

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

Steiner, M. (1978) "Mathematical Explanation," Philosophical Studies 34: 135 – 151.

Woodward, J. (1997a) "Explanation, Invariance and Intervention," PSA 1996, vol. 2, S26- 41.

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

Woodward, J., and C. Hitchcock (forthcoming) "Explanatory Generalizations, Part I: A Counterfactual Account," Noûs.

NOTES

* We would like to thank Nancy Cartwright, Malcolm Forster, Alan Hájek, Dan Hausman, Richard Healey, Paul Humphreys, and Judea Pearl for helpful comments. Woodward 's contribution to this paper was supported in part by the National Science Foundation (SBR-9320097).

1 Woodward and Hitchcock (forthcoming) contains a detailed comparison with Hempel's D-N model (Hempel and Oppenheim 1948, Hempel 1965a).

2 As we note in EG1, requiring that an intervention on X affect Y, if at all, only through the effect of the intervention on X is not tantamount to requiring that the intervention on X does affect Y.

3 For additional discussion, see Woodward (2000, sections 8 and 9).

4 Note that it will not help to explicitly incorporate various ceteris paribus conditions — e.g., the mass and radius of the earth — as antecedents in order to render Galileo's law exceptionless. The resulting law would still say nothing about what would happen if these conditions were changed, and hence would not be invariant under testing interventions on the relevant variables.

5 We are grateful to Richard Healey for a helpful discussion of this issue.

6 Lewis contrasts explaining an 'event' with explaining other types of phenomena, such as generalizations, or capacities of systems; these explanations need not be causal. Since this distinction will not be central to our discussion, we will typically drop explicit reference to explanation of events.

7 Hitchcock (1995) levels a similar objection against Salmon (1984).