Brit. J. Phil. Sci. **51** (2000), 197–254

Explanation and Invariance in the Special Sciences

James Woodward

ABSTRACT

This paper describes an alternative to the common view that explanation in the special sciences involves subsumption under laws. According to this alternative, whether or not a generalization can be used to explain has to do with whether it is invariant rather than with whether it is lawful. A generalization is invariant if it is stable or robust in the sense that it would continue to hold under a relevant class of changes. Unlike lawfulness, invariance comes in degrees and has other features that are well suited to capture the characteristics of explanatory generalizations in the special sciences. For example, a generalization can be invariant even if it has exceptions or holds only over a limited spatio-temporal interval. The notion of invariance can be used to resolve a number of dilemmas that arise in standard treatments of explanatory generalizations in the special sciences.

1 Introduction

A central problem in the philosophy of the special sciences concerns the nature and status of explanatory generalizations in those disciplines. Many philosophers are committed to a *nomothetic* conception of explanation according to which all successful explanations must appeal to laws. The standard assumption about laws is that they are exceptionless generalizations meeting various

other familiar conditions—they must contain only qualitative predicates, support counterfactuals, and so on. Together these assumptions generate a dilemma. On the one hand, most of us believe that the special sciences sometimes succeed in providing explanations. On the other, it looks as though most generalizations in the special sciences fail to conform to the standard criteria for lawhood—for example, they are not exceptionless and hold at best over limited domains or spatio-temporal intervals. The usual strategy for resolving this difficulty is to argue, despite all appearances to the contrary, that explanatory generalizations in the special sciences do meet, or somehow serve as stand-ins for generalizations that meet, the standard conditions for lawhood. The appeal of this strategy lies not in its inherent plausibility but rather in the difficulty of formulating a defensible alternative to the nomothetic conception of explanation.

This paper explores a new way out of this dilemma. It will argue that we need to rethink both the nomothetic conception of explanation and the standard conditions for lawhood, at least in so far as these are taken to provide criteria for distinguishing explanatory from unexplanatory generalizations. The standard framework suggests that there are just two, mutually exclusive possibilities: either a generalization is a law or else it is purely accidental. Most explanatory generalizations in the special sciences do not fit comfortably into either of these two categories. What we need is a new way of thinking about generalizations and the role that they play in explanation that allows us to recognize intermediate possibilities besides laws and accidents and to distinguish among these with respect to their degree or kind of contingency. This account should also allows us to understand how a generalization can play an explanatory role even though it holds only within a certain domain or over a limited spatio-temporal interval and has exceptions outside of these.

The alternative account I will propose rests on several key ideas. The first is a claim about explanation: explanatory relationships are relationships that *in principle* can be used for manipulation and control in the sense that they tell us how certain (explanandum) variables would change if other (explanans) variables were to be changed or manipulated. The qualification 'in principle' means that what matters for the purposes of explanation is not whether the manipulation in question can actually be carried out but rather whether the putative explanatory relationship correctly describes what would happen on the (possibly counterfactual) supposition that the manipulation is carried out.

Second, given this conception of explanation, it follows that whether or not a generalization can be used to explain has to do with whether it is *invariant* rather than with whether it is lawful. A generalization is invariant if (i) it is, in a sense I will try to make more precise below, change-relating and (ii) it is stable or robust in the sense that it would continue to hold under a special sort of change called an *intervention*. When invariance is so characterized, some laws

turn out not to be invariant because they are not change-relating. Hence some laws are not explanatory. More importantly, there are many examples of invariant relationships that are not laws. Appeal to laws is thus neither sufficient nor necessary for successful explanation. In contrast to the standard notion of lawfulness, the notion of invariance is well suited to capturing the distinctive characteristics of explanatory generalizations in the special sciences. A generalization can be invariant within a certain domain even though it has exceptions outside that domain. Moreover, unlike lawfulness, invariance comes in gradations or degrees.

As remarked above, to characterize invariance we need the notion of an intervention which we can think of as an idealized experimental manipulation. An intervention is an exogenous causal process that changes some variable of interest X in such a way that any change in some second variable Y occurs entirely as the result of the change in X. On the conception I will be defending, we may think of explanation as having to do not with subsumption under laws but rather with the exhibition of patterns of counterfactual dependence of a special sort, involving *active* counterfactuals—counterfactuals the antecedents of which are made true by interventions. Only invariant generalizations will support active counterfactuals—hence the connection between explanation and invariance.

Before turning to details, a few general remarks about the scope of this paper are in order. First, the account that follows is intended as an account of causal explanation and attempts to capture a notion of explanation according to which to explain is to cite causes. It is not intended as an account of non-causal forms of explanation, if there are such. Second, I will focus on cases in which generalizations are used to explain explananda that are at least implicitly general or repeatable. I will ignore issues about the structure of singular causal explanations in which what is explained is a particular event and problems of explanatory (or causal) overdetermination or pre-emption that arise in connection with such explanations.

2 Interventions

I begin with the notion of an intervention. Heuristically, one may think of an intervention as an idealization of an experimental manipulation carried out on some variable X for the purpose of ascertaining whether changes in X are causally or nomologically related to changes in some other variable Y. However, as we shall see shortly, any process, whether or not it involves human beings or their activities, will qualify as an intervention as long as it has the right causal characteristics. The idea we want to capture is roughly this: an intervention on some variable X with respect to some second variable Y is a causal process that changes X in an appropriately exogenous way, so that if a

change in *Y* occurs, it occurs only in virtue of the change in *X* and not as a result of some other set of causal factors.

Suppose that one wants to know whether treatment with some drug is effective in producing recovery from a disease. We may represent the treatment received by an individual subject *i* by means of a binary variable *T* that takes one of two values 0 and 1 depending on whether *i* does or does not receive the drug. Similarly, recovery may be represented by means of a variable *R* taking values 0 and 1, depending on whether or not individuals with the disease recover. Intuitively what one wants to know is whether if some subject *i* who has not received the treatment and who suffers from the disease (for whom *T(i)* = 0 and *R(i)* = 0) were to be given the drug (i.e. if *T(i)* were to be changed to 1), *i* would recover or would be more likely to recover (whether *R(i)* would be changed to 1). Obviously, one cannot investigate this question by both giving the treatment to and withholding it from the same subject. However, one may employ a more indirect method: divide the subjects with the disease into a treatment and control group, intervene by giving the drug to the former and not the latter, and then observe the incidence of recovery in the two groups. The experimenter's interventions (which we may represent by means of an intervention variable *I*) will thus consist in the assignment of values of *T* to individual subjects. Obviously, these interventions must meet various further conditions if the experiment is to tell us anything about the efficacy of *T*. First, if the experimenter's interventions *I* are correlated with some other cause of recovery besides *T*, this may undermine the reliability of the experiment. This would happen, for example, if the patients in the treatment group were much healthier than those in the control group. However, it would be too strong to require that *I* (or *T*) be uncorrelated with all other causes of *R*. As long as *T* is efficacious, *I* and *T* will be correlated with other causes of *R* that are themselves caused by *I* or by *T*. For example, if treatment by the drug does cause recovery and does so by killing (*K*) a certain sort of bacterium, then it will be no threat to the validity of the experiment if the experimenters' interventions *I* are correlated *K*, even though *K* causally affects *R*. What we need to rule out is the possibility that there are causes of *R* that are correlated with *I* or caused by *I*, and that affect *R* independently of the $I \rightarrow T \rightarrow K \rightarrow R$ causal chain.

A third condition is that *I* should not directly affect recovery independently of *T* but only, if at all, through it. This means, among other things, that *I* must not be a common cause of both *T* and *R*. This condition would be violated if, for example, the subjects learn whether or not they have been assigned to the treatment group and the control group and this has a placebo effect—an effect on *R* that is independent of any effect of *T* itself on *R*. (Perhaps those in the treatment group are made more hopeful by the fact that they are in this group and those in the control group are discouraged.) In this case *I* directly affects *R*

independently of *T* and we will not be able to reach reliable conclusions about the effect of *T* on *R*.

Assembling these requirements together, we are led to the following characterization: Suppose that *I* is an intervention on (or manipulation of) the variable *X*, where *X* is some property possessed by the unit *i*, the intent being to assess some postulated relationship (G) according to which changes in *X* cause or explain changes in some other variable *Y* by observing whether the intervention on *X* produces the change in *Y* predicted by (G). Call the value of *X* possessed by *i* prior to the intervention x_0 and the value after the intervention x_1. Then *I* should have the following conjunction of features (M):

M1) *I* changes the value of *X* possessed by i from what it would have been in the absence of the intervention (i.e. $x_1 \neq x_0$) and this change in *X* is entirely due to *I*.

M2) The change in *X* produced by *I* is claimed by (G) to change the value of *Y*. That is, according to (G), the value, y_0, that *Y* takes when $X = x_0$, is different from the value, y_1, that *Y* takes when $X = x_1$.

M3) *I* changes *Y*, if at all, only through *X* and not directly or through some other route. That is, *I* does not directly cause *Y* and does not change any causes of *Y* that are distinct from *X* except, of course, for those causes of *Y*, if any, that are built into the *I*–*X*–*Y* connection itself; that is, except for (a) any causes of *Y* that are effects of *X* (i.e. variables that are causally between *X* and *Y*) and (b) any causes of *Y* that are between *I* and *X* and have no effect on *Y* independently of *X*. In addition, *I* does not change the causal relationships between *Y* and its other causes besides *X*. Moreover, a similar point holds for any cause *Z* of *I* itself—i.e. *Z* must change *Y*, if at all, only through *X* and not through some other route.

M4) *I* is not correlated with other causes of *Y* besides *X* (either *via* a common cause of *I* and *Y* or for some other reason) except for those falling under (M3a) and (M3b) above.

There are several features of this characterization that are worth noting. First, as advertised above, the characterization makes no essential reference to human activities or to what human beings can or can't do. A causal process that does not involve human beings at any point will qualify as an intervention as long as it meets conditions M1–4. Indeed, it is precisely this sort of possibility one has in mind when one talks about a 'natural experiment'.

A second issue concerns circularity. The characterization (M) employs causal language at a number of points—not only must the intervention *I* cause a change in the variable *X*, but *I* must not itself directly cause *Y*, must not be correlated with other causes of *Y* that are independent of the putative *I*→*X*→*Y* chain, and so on. Because the notion of an intervention is already a

causal notion, it follows that one cannot appeal to it to explain what it is for a relationship to be causal or nomological (or invariant) in terms of concepts that are themselves entirely non-causal or non-nomological. Nonetheless, it is important to understand that the characterization is not viciously circular in the sense that the characterization of an intervention on *X* with respect to *Y* itself makes reference to the presence or absence of a causal relationship between *X* and *Y*. Instead the characterization makes reference to *other* causal relationships—to the existence of a causal relationship between *I* and *X* and to the distribution of other possible causes of *Y* besides *X*. The characterization (M) thus fits with a non-reductive account of causal and nomological relationships and of how we infer to the existence of such relationships. The fundamental idea is that we can explain what it is for a relationship between *X* and *Y* to be causal or nomological (or invariant) by appealing to facts about *other* causal or nomological relationships involving *X* and *Y* and to non-causal correlational facts involving *X* and *Y*. That there is a coherent notion of an intervention to be captured, and that some explication of this notion that is not viciously circular must be possible, is strongly suggested by the fact that we do seem to sometimes find out whether a causal or nomological relationship exists between *X* and *Y* by manipulating *X* in an appropriate way and determining whether there is a correlated change in *Y*. This fact by itself seems to show that we must have some notion of a manipulation of *X* that would be suitable for finding out whether *X* is causally or nomological linked to *Y*, and that this notion can be characterized without presupposing that there is a causal or nomological relationship between *X* and *Y*. It is just this notion that (M) attempts to capture.

A third issue concerns clause (M1). This says that carrying out an intervention on *X* requires that there be a well-defined notion of changing the value of *X* possessed by some individual in such a way that the very same individual is caused by the intervention to possess a different value of *X*. One consequence of (M1) is that there is no well-defined notion of an intervention with respect to properties or magnitudes that, for logical or conceptual reasons, can only take one value. For example, if everything that exists is necessarily a physical object there is no well-defined notion of intervening to change whether something is a physical object. Even with respect to variables that can take more than one value, the notion of an intervention will not be well defined, if there is no well-defined notion of changing the values of that variable for a particular individual. For example, we might introduce a variable 'animal' which takes the values {trout, kitten, raven} but if, as I suspect, we have no coherent idea of what it is to change a raven into trout or kitten, there will be no well-defined notion of an intervention for this variable. This restriction on the notion of an intervention to variables for which there is a well-defined notion of change is both implicit in the notion of an intervention itself and also follows from our

guiding idea that explanatory relations are relations that can be used for manipulation and control. If there is no well-defined notion of changing the value of *X*, we cannot, even in principle, manipulate some other variable by changing *X*. Similarly, unless as (M2) requires, the contemplated intervention on *X* is, according to the generalization we are assessing, associated with a change in *Y*, this generalization will not tell us how intervening on *X* can be used to manipulate *Y*. As we will see below, both (M1) and (M2) have important consequences for the sorts of generalizations that can figure in explanations.

It will help to clarify the notion of an intervention if we consider an additional example. Suppose that, in a certain region, changes in atmospheric pressure (*A*) are a common cause and the only cause of the occurrence of storms (*S*) and of the reading (*B*) of a particular barometer and that there is no direct causal relationship between *B* and *S*. Imagine that we are ignorant of the causal structure of this system and wish to find out whether *B* directly causes *S* by changing *B* and ascertaining whether there is a corresponding change in *S*. It is clear that certain ways of changing *B* are inappropriate for this purpose. If we change *B* by changing *A*, or by means of some causal process that is perfectly correlated with changes in *A*, then *S* will also change, but this will not establish that there is a causal relationship between *B* and *S*. Similarly, if we change *B* via some process that directly affects *S*. None of these ways of changing *B* will qualify as interventions on *B* for the purpose of ascertaining whether there is a causal relationship between *B* and *S*—they run afoul, respectively, of clauses (M3) and (M2) in (M).

By contrast, suppose that we employ a random number generator which is causally independent of *A* and, depending just on the output of this device, repeatedly physically fix the barometer reading at different values by moving the dial to either a high or low reading and driving a nail through it. Suppose that this procedure results in repeated settings of the dial that are uncorrelated with *A*. If—as it seem reasonable to believe—this procedure satisfies the other conditions in (M), repetitions of it will count as interventions on *B* with respect to *S*. This illustrates the sense in which interventions involve *exogenous* changes—such changes in the *A*–*B*–*S* system are exogenous in the sense that they do not operate through *A* or through processes correlated with *A*. Such changes break or disrupt the previously existing endogenous causal relationship between *A* and *B* since the state of *B* is now set by the intervention, independently of *A*. When the barometer reading is changed in this way, what we expect of course is that the previous association or correlation between *B* and *S* will break down or disappear (that it will be non-invariant) and hence that the relationship between *B* and *S* will fail to qualify as a causal or nomological.

Given that our ultimate interest is in characterizing what it is for there to be a causal or explanatory relationship between *X* and *Y*, why not drop any

reference to interventions on *X* and simply talk directly about the nature of the relationship between *X* and *Y*? As I see it, invoking the notion of an intervention (and the closely connected notions of invariance and active counterfactual dependence) has several important advantages. First, to say simply that for *X* to cause or figure in an explanation of *Y*, there must be a causal or explanatory relationship between *X* and *Y* is to move in a very small and completely unilluminating circle. What we want is some sort of independent purchase on what it is for the *X*-*Y* relationship to be causal or explanatory and this is what talk of the behavior of *Y* under interventions on *X* is intended to provide. (That this genuinely provides an independent purchase is shown by the fact that we can use the behavior of *Y* under interventions on *X* to discover whether there is causal or explanatory relationship between *X* and *Y*—something that, as noted above, would be impossible if to recognize such behavior we already had to know whether there is a causal relationship between *X* and *Y*.) As the example in the immediately preceding paragraphs illustrates, some variables (e.g. *B* and *S*) will be correlated (and in this sense the value of one will depend on the value of the other) even though neither causes nor figures in an explanation of the other while other variables (e.g. *A* and *S*) will be correlated and causally and explanatorily related. Invoking the notion of an intervention gives us a way of distinguishing between those sorts of correlations and dependencies that reflect causal and explanatory relations and those that do not. Of course what matters for whether *X* causes or explains *Y* is the 'intrinsic' character of the *X*–*Y* relationship but the attractiveness of the notion of an intervention is precisely that it provides an extrinsic way of picking out or specifying this intrinsic feature.

A second advantage of the characterization offered above is that it makes it the epistemological role of experimentation in establishing causal and explanatory relationships particularly transparent. While it is no part of my view that the only way we can establish whether there is a causal relationship between *X* and *Y* is by actually carrying out an intervention on *X* and observing the response of *Y*, it is uncontroversial that this is a particularly effective way of establishing causal relationships. The account described above explains why this should be so, since it builds claims about behavior under interventions into the very content of causal claims. The epistemological role of experimentation is considerably less clear on alternative accounts of causation and explanation.[1]

[1] An anonymous referee worries that my appeal to interventions confuses the epistemology and metaphysics of explanation: seeing if *Y* changes under an intervention on *X* is a good way of finding out whether *X* causes or explains Y but does not capture what it is for *X* to cause *Y*. This worry would be well founded if we had available some independent alternative (and non-trivial) characterization of the conditions that must be met for it to be true that *X* causes *Y* for then the behavior of *Y* under interventions on *X* would be just a test for whether those conditions obtain and not a characterization of those conditions themselves. However, I deny that we have any such alternative characterization. On my view, for *Y* to change under an appropriate intervention on *X*

3 Invariance

Once we have the notion of an intervention, we can use it to characterize more precisely the notion of invariance under interventions, which I take to be the key feature that a generalization must possess if it is to play an explanatory role or to describe a causal or nomological relationship. The general idea of invariance is this: a generalization describing a relationship between two or more variables is invariant if it would continue to hold—would remain stable or unchanged—as various other conditions change. The set or range of changes over which a relationship or generalization is invariant is its *domain* of invariance. As we will see in more detail below, invariance is a relative matter—typically a relationship will be invariant with respect to a certain range of changes but not with respect to other changes.

It is useful to distinguish two sorts of changes that are relevant to the assessment of invariance. First, there are changes in what we would intuitively regard as the background conditions to some generalization—changes that affect other variables besides those that figure in the generalization itself. For example, in the case of a system of masses conforming to the gravitational inverse square law (1) $F = Gm_1 m_2/r^2$, changes in the position or velocity of the system as a whole which do not change the relative positions of the masses will count as a changes in background conditions, as will a change in the color of the masses or their electrical charge, or the Dow–Jones Industrial average. The inverse square law is invariant under changes in all these background conditions.

Second, there are changes in those variables that figure explicitly in the generalization itself—for example, in the case of (1), mass and distance. An important subclass of such changes are changes that result from an intervention (in the sense specified in section **1**) on the variables figuring in the generalization. The gravitational inverse square law is invariant not just under changes in background conditions but also under a wide range of interventions that change the distances between gravitating masses or the magnitudes of the

just is what it is for *X* to cause *Y*. For example, when we attempt to infer to causal conclusions on the basis of non-experimental data, as is the case with the structural equation methods discussed in Section **12**, what we are trying to do is to use such data to infer what would happen if a hypothetical experiment were to be carried out. The account I present is thus intended as a characterization of the metaphysics of causation (or of the truth conditions for causal claims) and not just as an account of one way of finding out about causal relationships.

It is also worth noting in this connection that theories that attempt to provide an account of the metaphysics of causation that does not appeal to the notion of behavior under an intervention have their own costs—they lack any plausible connection with the epistemology of causation. For example, if, as Salmon ([1984]) holds, the metaphysics of causation involves spatio-temporally continuous processes that transfer energy and momentum, then we require some account, which Salmon does not provide, of why experiments are a good way of finding out about causal processes so conceived and which experiments are the appropriate ones to consider. The well-known criticisms of Salmon in Kitcher ([1989)] exploit this point. The intervention based account avoids such difficulties.

masses themselves. We will see later that this is crucial to its explanatory status. I will say that a generalization is invariant *simpliciter* if and only if (i) the notion of an intervention is applicable to or well-defined in connection with the variables figuring in the generalization (see below) and (ii) the generalization is invariant under at least some interventions on such variables. In other words, for a generalization to count as invariant there must exist some interventions (satisfying the conditions (M1)–(M4)) for variables figuring in the relationship under which it is invariant. To count as invariant it is *not* required that a generalization be invariant under all interventions. For brevity, I will often speak of a generalization as 'invariant under interventions' as shorthand for 'invariant under some interventions on variables explicitly figuring in the generalization'.

The generalization (1) is naturally regarded as a generalization that relates changes—it describes how changing the magnitudes of two masses or the distance between them will change the gravitational force they exert on each other. Other generalizations, including some we may regard as laws, are not naturally interpreted as true descriptions of relationships between changes. Such generalizations fall into several categories. First, there are generalizations that, to put it loosely, do not tell us how to produce certain changes but rather that they are impossible. Consider the generalization (2) 'No material object can be accelerated from a velocity less than that of light to a velocity greater than that of light'. This generalization does not tell us, as (1) does, how changes in one set of variables will produce changes in another set of variables. Instead, it rather tells us, in effect, that there are no physically possible changes that will produce a change from subluminal to superluminal velocities.

As explained above, it is a consequence of clause (M1) in (M) that the notion of an intervention (and hence the notion of invariance under interventions) is well defined only for change-relating generalizations. If a generalization (G) relates some variable *X* to some other variable *Y* but it makes no sense to speak of changing the value of *X* for some individual to a different value, then there will be no well-defined notion of an intervention for that variable. The notion of an intervention does not seem to be well defined in connection with (2). In so far as the notion of a material object in (2) contrasts with anything at all, it presumably contrasts with the notion of a pseudo-process in the sense of Salmon ([1984]). It is arguable that there is no well-defined notion of changing a pseudo-process into a material object or *vice versa* and if so, (2) will not be invariant in the sense described above. Moreover, even if the latter change (from material object to pseudo-process) were well defined, it would fail to satisfy clause (M2) in the characterization of an intervention since (2) does not predict that the object so altered would change its velocity from sub- to superluminal. That is, what (2) claims is simply that material objects cannot undergo a certain change—it says nothing at all about the behavior

of non-material objects, if there are any. There is, to be sure, another, intuitive sense in which is (2) highly invariant—there are no physically possible changes (nothing that we or the rest of nature might do) that will disrupt it. It is at least in part for this reason that we think of it as a law. However, it is not invariant under interventions on the variables figuring in the relationship in the sense defined above and hence is not explanatory (*cf.* Section **4**, fn. 3).

A second possibility is illustrated by the generalization (3) 'All men who take birth control pills regularly fail to get pregnant' (*cf.* Salmon [1971]). There are at least two possible ways of understanding this generalization. First, it may be understood as a generalization that does not even purport to be change-relating. Interpreted in this way, (3) does not claim that changes in whether or not men take birth control pills are correlated with changes in whether or not they become pregnant but only that male pill-takers do not get pregnant. It says nothing about whether males who do not take birth control pills will get pregnant. Under this interpretation, (3) is true but there is no well-defined notion of an intervention associated with it and it fails to be invariant for this reason. Second, (3) may be interpreted as claiming that the correlation described above does hold. Under this interpretation, while there is a well-defined notion of intervention associated with (3) (one can intervene to change men who are pill-takers to non-pill-takers and *vice versa*), (3) is false, since the claimed correlation fails to hold: whether or not a male takes birth control pills is not correlated with whether he becomes pregnant. Intuitively, (3) cannot be used to explain why some particular man fails to get pregnant, because taking birth control pills is irrelevant to whether a man becomes pregnant. We will see in Section **4** how the fact that (3) is not an invariant change-relating generalization underlies this judgment of irrelevance.

When a generalization like (1) that relates changes is invariant under interventions on variables figuring in the relationship, it describes a relationship that is hypothetically exploitable for purposes of manipulation and control—hypothetically exploitable in the sense that although it may not always be possible, as a practical matter, to intervene to change the values of the quantities described by the variables that figure in the generalization, we can nonetheless think of the generalization as telling us that *if* it were possible to change those values, one could use them to change others. Thus, for example, because the gravitational inverse square law is invariant under a range of interventions that change mass and distance, it tells us how, if we or some natural process were to manipulate these quantities in some system of gravitating masses, the gravitational force they exert would change in a systematic way. Similarly, because (4), the ideal gas law $PV = nRT$, is invariant under a range of interventions that change temperature, it correctly describes how, by manipulating the temperature of a gas and holding its volume constant, one could change its pressure.

Why does it matter whether a generalization is invariant under (at least some) interventions on the variables figuring in the relationship as opposed to merely being invariant under some changes in background conditions? The reason is that any generalization, no matter how 'accidental', non-lawful, non-causal or unexplanatory, will be stable or will continue to hold under some changes in background conditions—for example, under changes in conditions that are causally independent of the factors related by the generalization. Thus, special circumstances aside, a paradigmatic accidental generalization like

(5) All the coins in Bill Clinton's pocket on January 8, 1999 are dimes.

will continue to describe correctly the contents of Clinton's pockets on this date under many possible changes in background conditions. For example, presumably (5) will be stable under changes in the position of Mars, the leadership of China, or the barometric pressure in Paris, and so on. Similarly, the generalization (6), describing the relationship between the barometer reading *B* and the onset of a storm *S* considered in Section **2**, will be stable under many changes in background conditions.

The feature that distinguishes generalizations like (5) and (6) from generalizations like (1) and (4) is that (5) and (6) are not invariant under interventions on the variables that explicitly figure in them and do not describe relationships that tell us how by manipulating one variable we may change or manipulate another—they do not describe relationships that are hypothetically exploitable for purposes of manipulation and control. To see this consider what an intervention would involve in the case of (5). To apply the characterization (M) to (5) we must interpret (5) as a change-relating generalization—i.e. as claiming that changing whether or not a coin is located in Clinton's pocket changes whether or not it is a dime. More specifically, think of *X* as a variable that measures whether or not a coin is located within Clinton's pocket and *Y* as a variable that measures whether or not it is a dime. According to clause (M2) in (M) for the introduction of a coin into Clinton's pocket to qualify as intervention, the coin must be such that (G) claims that its value would change if its location were changed from being outside Clinton's pocket to being inside. The introduction of a dime into Clinton's pocket will not meet this condition but the introduction of a penny would. However, (5) is plainly not invariant under such interventions—the introduction of non-dimes into Clinton's pocket will not transform them into dimes. More generally, (5) is not invariant under any interventions on the variable *X*. A similar point holds for (6). As observed in Section **2**, (6) is not invariant under interventions that consist in manipulating the barometer dial. While (4), the ideal gas law, tells us how we can make the pressure or volume of a gas change by changing its temperature, we cannot change non-dimes into dimes by introducing them into Clinton's pocket. And we cannot alter whether or not storms occur by fiddling

with barometer dials. I will suggest below that this difference between (5) and (6), on the one hand, and (1) and (4), on the other, is crucial to their explanatory status.

4 Explanation

I suggested above that explanation requires appeal to invariant generalizations. In this section I want to sketch an account of explanation that supports this suggestion.

Consider a gas enclosed in a rigid container of volume V^* which under goes a temperature increase to T^* in virtue of being connected to a heat source. If we want to provide a simple but none the less genuine explanation of (7), why the pressure of the gas increases to P^*, it seems relevant to cite the new temperature T^*, the constant volume V^* and the ideal gas law (4). According to an account of explanation which I have defended in more detail elsewhere (Woodward [1979], [1984], [1997]), this information is explanatory because it can be used to answer a range of counterfactual or what-if-things-had-been-different questions about the explanandum (7). What I mean by this is that the generalization (4) can be used, in conjunction with information about the 'initial conditions' of the gas (the fact that it has temperature T^* and volume V^*) to show how the explanandum (7) would change, if these initial conditions were to change in various ways. That is, not only can (4) be used, in conjunction with information about the initial conditions of the gas to show that the explanandum (7) 'was to be expected', as the traditional DN model demands, but (4) can also be used to tells us how the pressure of the gas would *change*—how the pressure would have been different—if the temperature had instead increased or decreased to a different value T^{**} or if the volume of the container had changed to a different value V^{**}. In this way the explanation in terms of (4) locates the explanandum (7) within a space of alternative possibilities (other possible values for the pressure that might have occurred) and shows us how which of these alternative possibilities is realized systematically depends on the initial temperature and volume of the gas. In seeing how the actual pressure P^* would have been different, had the actual temperature T^* and volume V^* of the gas been different, we see in detail how the pressure depends on these factors and how they are explanatory relevant to the pressure. In short, we can think of the explanation that appeals to (4) as exhibiting a systematic pattern of counterfactual dependence of the pressure of the gas on its temperature and volume. The exhibition of such a pattern is at the heart of successful explanation.

Consider another example. The gravitational inverse square law allows us to see how the gravitational force between two or more objects would have been different had the distances between these objects or their masses had been

different in various ways. In combination with the Newtonian laws of motion, this information allows us to see that, given an object with a certain mass, initial position and velocity, it will follow a certain trajectory under the gravitational influence of a second object. However, these laws also enable us to see how this trajectory would change given changes in these initial conditions or in the mass of the attracting object. For example, we can use the inverse square law and the equations of motion to see how under certain conditions the first object will follow an elliptical orbit about the second, how under other conditions it will spiral into the second and so on. In this way we see how the actual trajectory depends on these factors and in seeing this we come to understand why the actual trajectory took the form it did.

The intuitive attractiveness of the idea that explanation has to do with the exhibition of systematic patterns of counterfactual dependence is further reinforced by the fact that we usually think of explanation as having to do with the exhibition of causal relationships and it seems undeniable that there is a close connection between causal relationships and counterfactuals. If explanations cite causes, it seems very plausible that some form of counterfactual theory of explanation must be correct.

However, familiar difficulties with counterfactual theories of causation also remind us that there is an obvious objection to such a proposal: there are relationships of counterfactual dependence which are not causal or explanatory. For example, assuming that the barometer–storm–atmospheric pressure system of Section **2** operates deterministically and that *A* is the only cause of *B* and *S*, it looks as though there is a perfectly good sense in which it is true that

(8) If the reading *B* of the barometer were falling, a storm would occur

and also true that

(9) If the reading *B* were rising, there would be no storm

However, despite the fact that *S* counterfactually depends on *B*, *B* doesn't cause *S* and one can't appeal to *B* to explain *S*.

We can deal with this difficulty by appealing to the ideas introduced in Section **2**. As explained earlier, the correlation between *B* and *S* is not invariant under interventions on *B*. While there is (arguably) an interpretation of the counterfactuals (8) and (9) according to which they are true, there is also natural interpretation of the counterfactual

(10) If an intervention were to occur which lowers (increases) *B*, then the storm would occur (not occur)

according to which it is false. This is the interpretation we adopt when we take (10) to be claiming that an intervention on *B* would be a way of controlling or manipulating or changing whether or not a storm occurs and that the previously

obtaining correlation between *B* and *S* would be invariant under such interventions. Let us say that when (10) is interpreted in this way, it is an *active* counterfactual. This reading contrasts with the *passive* interpretation of the counterfactuals (8) and (9) appealed to above which carries with it no such claim about would happen to *S* under interventions on *B*. My claim is that the kind of counterfactual dependence that matters for successful explanation is active counterfactual dependence.[2] Put differently, a successful explanation should appeal to factors or variables such that interventions on those factors will be systematically associated with corresponding changes in its explanandum. An explanation of the pressure of a gas in terms of its pressure and volume meets this requirement—it supports or is associated with active counterfactuals while a purported explanation of the occurrence of a storm that appeals to the correlation between *B* and *S* does not.

On this understanding of what an explanation does, the connection between explanation and invariance should be transparent. It is only if a generalization is invariant under some range of interventions and changes that we can appeal to it to answer what-if-things-had-been-different questions about what would happen under these interventions and changes. For example, if the ideal gas law systematically broke down under interventions that change temperature and volume, then we could not appeal to it to answer questions about how the pressure of a gas would change under such changes. Similarly it is because the gravitational inverse square law is invariant under interventions that change the distances between various objects and their masses that we can use this generalization to show how the gravitational forces exerted by those objects would change if their masses and the distances between them would change.

I can further clarify this account of explanation by means of an additional comparison with the most familiar rival account—the DN model defended in Hempel ([1965]). In the examples described above, we are shown, just as the DN model demands, that an explanandum (e.g. (7)) is deducible from a law

[2] The contrast between active and passive counterfactuals roughly corresponds to the contrast between non-backtracking and backtracking counterfactuals in David Lewis's sense ([1973b]). A counterfactual whose antecedent is made true by an intervention will behave in roughly the same way as a counterfactual whose antecedent is made true by a 'small' Lewisian miracle. However, this correspondence is only rough and not exact. Even if we put aside the fact that Lewis's theory focuses only on causal relationships between particular events, the intervention-based approach that I advocate differs from Lewis's in some of the specific judgments that it reaches about which relationships are causal or explanatory—see Hitchcock and Woodward ([unpublished manuscript]) for additional discussion. In addition, Lewis' theory is reductionist in aspiration: he hopes to explain the notion of causation in terms of a more general notion of counterfactual dependence that does not itself presuppose causal notions. By contrast, I do not think that such a reduction is possible. While the matter deserves more detailed discussion than I can give it here, my view is that , given that *c* causes *e*, which counterfactual claims involving *c* and *e* are true will depend on which other causal claims involving other variables besides *c* and *e* are true in the situation under discussion. For example, it will depend on whether other causes of *e* besides *c* are present. In my view, reference to these other causal claims cannot be eliminated in favor of purely noncausal counterfactual talk. As explained above, (M) recognizes this.

(e.g. (4)) and a statement of initial conditions. However, we are also shown something in addition to this—namely how (4) can be used to answer a set of what-if-things-had-been-different questions about (7). This represents an independent condition or constraint, that has no counterpart in the DN model and does not follow just from the DN requirement that (7) be derivable from a law and a statement of initial conditions. One way of bringing this out is to remind ourselves of a familiar counter-example to the DN model, drawn from Salmon ([1971]). Consider the derivation:

(Ex. 11)
(L_{11}) All men who take birth control pills regularly fail to get pregnant.
(C_{11}) Mr. Jones is a man who takes birth control pills regularly.
(E_{11}) Mr. Jones fails to get pregnant.

If we agree that (L_{11}) is a law, (Ex. 11) meets the conditions for successful DN explanation. None the less it is a defective explanation. The theory I have proposed explains why in a natural way: the condition cited in the explanans of (Ex. 11) is not such that changes in it produced by interventions would lead to changes in the outcome being explained. A change in whether Jones takes birth control pills will lead to no change in whether or not he gets pregnant. In consequence, (Ex. 11) fails to satisfactorily answer a what-if-things-had-been-different question about its explanandum: it fails to correctly identify the conditions under which an outcome different from (E_{11}) would occur and, indeed, wrongly suggests that the condition cited (taking birth control pills) is a condition such that changes in it would lead to changes in whether Jones gets pregnant when in fact this is not true. This failure is reflected in our judgment that taking birth control pills is explanatorily irrelevant to whether Jones gets pregnant. Put differently, what (Ex. 11) shows is that a derivation can cite a nomologically sufficient condition for an explanandum and yet fail to answer a what-if-things-had-been-different question about it and hence fail to explain it. Hence it shows how the account that I have provided differs from accounts that take explanation to be just a matter of derivation from a law.

Consider, by way of contrast (Ex. 12):

(L_{12}) All women who meet condition *K* (*K* has to do with whether the woman is fertile, has been having intercourse regularly and so forth) and who take birth control pills regularly will not get pregnant and furthermore all women who meet condition *K* and do not take birth control pills regularly will get pregnant.
(C_{12}) Ms. Jones is a woman who meets condition *K* and has been having intercourse regularly.
(M_{12}) Ms. Jones does not get pregnant.

Here, of course, we have considerably more inclination to say that at least a

crude explanation of (M_{12}) has been provided. On the account that I advocate this difference between (Ex. 11) and (Ex. 12) is a reflection of the fact that the latter, but not the former, satisfies the what-if-things-had-been-different condition on explanation. The condition *K* cited in (Ex. 12) is such that changes in it would lead to changes in the outcome being explained—if Ms. Jones stops taking birth control pills, is fertile and has intercourse, she will or at least may get pregnant; if she fails to take the pills but also doesn't have intercourse she will not get pregnant and so on. (Ex. 12) thus draws our attention to a systematic pattern of active counterfactual dependency of changes in its explanandum (E_{12}) on changes in its explanans. Unlike (Ex. 11), (Ex. 12) does locate its explanandum within a range of possible alternatives and shows, at least in a crude way, the range of conditions under which this explanandum would hold and what sorts of changes in those conditions would instead lead to one of these alternatives. In doing this (Ex. 12) shows us as how the conditions cited in its explanans make a difference for, or are explanatorily relevant to its explanandum. To put the point a bit differently, the contrast between (Ex. 11) and (Ex. 12) shows that explanatory relevance—the key feature that is lacking in (Ex.11) but present in (Ex. 12)—is just a matter of the holding of the right sort of pattern of active counterfactual dependence between explanans and explanandum: to a first approximation, *S* is explanatorily irrelevant to *M* if *M* would hold both if *S* were to hold and if *S* were not to hold when these counterfactuals are interpreted actively.[3]

There is a second respect in which the account I am recommending differs from the traditional DN account. The DN model requires that every explanation must appeal to at least one law. Since laws (or more precisely laws that connect changes and do not contain irrelevancies in the sense specified in Section **3**), in virtue of their invariance characteristics and the support they provide for active counterfactuals, provide information relevant to answering what-if-things-had-been-different questions, my account agrees with the DN model in holding that laws play a crucial role in (at least) some explanations. That is, one way, illustrated by the examples described above, in which an explanation can satisfy the requirement that it provide answers to a range of what-if-things-had-been-different questions is by appealing to a law. However, in taking invariance and support for active counterfactuals rather than lawfulness *per se* to be crucial to successful explanation we open up an intriguing

[3] This is by no means the only kind of case in which the DN model and the what-if-things-had-been-different account of explanation diverge. For another sort of illustration consider laws that do not relate changes in the sense described in Section **2**. These can be used to explain according to the DN model, but not according to the what-if-things-had-been-different account. For example, according to the DN model, one can explain why some particular massive particle moves at a subluminal velocity by appealing to the law (2) which states that all massive particles behave in this way. The model I favor denies this since the proposed explanation does not tell us anything about the conditions under which massive particles will not behave in this way. Another sort of case in which the model I favor and the DN model differ is described in fn. 4.

possibility that is not available within the DN or other purely nomothetic frameworks: the possibility that there are generalizations that are invariant and that can be used to answer a range of what-if-things-had-been-different questions and that hence are explanatory, even though we may not wish to regard them as laws and even though they lack many of the features traditionally assigned to laws by philosophers. In Sections **12** and **13**, I will argue that many explanatory generalizations in the special sciences have exactly this character—they are invariant generalizations that are not naturally regarded as laws. To put the point a bit differently, the account provided above allows us to reject the assumption, accepted by many if not most philosophical commentators, that if the generalizations of the special sciences are genuinely explanatory, they must either be or be closely associated in some way with laws of nature. It thus allows us to avoid the various puzzles and difficulties that nomothetic conceptions of explanation encounter when we attempt to apply them to the special sciences.[4]

5 Degrees of invariance

My argument so far has been that generalizations like (2) and (3), in contrast to (1) and (4) are not invariant under (any) interventions (that is, on the variables that explicitly figure in those generalizations) at all—they are, as I shall say, non-invariant and hence non-explanatory. However, as already intimated, invariance is not an all or nothing matter. Most generalizations that are invariant under some interventions and changes are not invariant under others. As we shall see shortly, we may legitimately speak of some generalizations as more invariant than others—more invariant in the sense that they are invariant under a larger or more important set of changes and interventions than other generalizations. Moreover, there is a connection between range of invariance and explanatory depth—generalizations that are invariant under a larger and more important set of changes often can be used to provide better explanations and are valued in science for just this reason. The picture that I will be defending is thus one in which there is both a threshold—some generalizations fail to qualify as invariant or explanatory at all because they are not invariant under any interventions on the variables that explicitly figure in the generalization—and above this threshold a notion of invariance that admits distinctions or gradations of various sorts. This picture corresponds to

[4] The examples of explanations considered so far and indeed all of the examples considered below involve deductive arguments. However, it is not part of the account I am proposing that all explanation must be deductive. The exhibition of a deductive structure is just one way in which an explanation can answer what-if-things-had-been-different questions. So-called singular causal explanations like 'The short circuit caused the fire' are explanatory because they answer what-if-things-had-been-different questions even though they are not deductive explanations—see Woodward ([1984]) for additional discussion. This is another respect in which the account that I am proposing differs from the DN model.

how, intuitively, we seem to think about explanatory (or causal or nomological relationships). Some relationships—e.g. the relationship described by (3)—are not causal or explanatory at all, but among those that are, some may be used to provide deeper or more perspicuous explanations than others. This represents just one of many points at which the invariance based account that I will be defending contrasts with more traditional frameworks for thinking about laws and their role in explanation. The traditional frameworks suggest a dichotomy: that either a generalization is a law or else it is purely accidental. Moreover, it is assumed that the boundary between laws and non-laws coincides with the boundary between those generalizations that can be used to explain and those that cannot. The invariance-based account rejects both of these ideas.

The ideas introduced in the previous paragraph—that generalizations may differ in the range of changes or interventions over which they are invariant and that these differences are connected to differences in their explanatory status—are familiar themes in the econometrics literature. They are illustrated and endorsed by Tygre Haavelmo, one of the founding figures of econometrics, in a well-known passage from his monograph 'The Probability Approach in Econometrics' ([1944]). In this passage, Haavelmo introduces a notion which he calls autonomy but is really just another name for what we have been calling invariance. He writes:

> If we should make a series of speed tests with an automobile, driving on a flat, dry road, we might be able to establish a very accurate functional relationship between the pressure on the gas throttle (or the distance of the gas pedal from the bottom of the car) and the corresponding maximum speed of the car. And the knowledge of this relationship might be sufficient to operate the car at a prescribed speed. But if a man did not know anything about automobiles, and he wanted to understand how they work, we should not advise him to spend time and effort in measuring a relationship like that. Why? Because (1) such a relation leaves the whole inner mechanism of a car in complete mystery, and (2) such a relation might break down at any time, as soon as there is some disorder or change in any working part of the car. We say that such a relation has very little autonomy, because its existence depends upon the simultaneous fulfillment of a great many other relations, some of which are of a transitory nature. On the other hand, the general laws of thermodynamics, the dynamics of function, etc., etc., are highly autonomous relations with respect to the automobile mechanism, because these relations describe the functioning of some parts of the mechanism *irrespective* of what happens in some other parts ([1944], pp. 27–8).

Haavelmo then suggests the following, more formal characterization of autonomy:

> Suppose that it would be possible to define a *class S, of structures*, such that *one member or another* of this class would, approximately, describe

> economic reality in *any practically conceivable situation*. And suppose that we define some non-negative measure of the 'size' (or the 'importance' or 'credibility') of any subclass, *W* in *S* including itself, such that, if a subclass contains completely another subclass, the measure of the former is greater than, or at least equal to, that of the latter, and such that the measure of *S* is positive. Now consider a particular subclass (of *S*), containing all those—and only those—structures that satisfy a particular relation '*A*' is autonomous with respect to the subclass of structures W_A. And we say '*A*' has a degree of autonomy which is the greater the larger the 'size' of W_A as compared with that of *S* ([1944], pp. 28–9).

Although this characterization is far from completely transparent (among other things, Haavelmo does not tell us how to go about determining the 'size' or 'importance' of *W*—matters which we will address below), the underlying idea is perhaps clear enough. In the most general sense the degree of autonomy of a relationship has to do with whether it would remain stable or invariant under various possible changes. (As we have argued, if, like Haavelmo, we wish to use this idea to distinguish between those relationships to which we can appeal to explain and those that cannot be so used, we need to include, among the changes over which we demand that a relationship be autonomous, those that correspond to interventions on the variables figuring in the relationship.) The larger the class of changes under which the relation would remain invariant—the more structures in *W* compatible with relation—the greater its degree of autonomy. Haavelmo suggests that physical laws such as the laws of thermodynamics and fundamental engineering principles such as those governing the internal mechanism of the car will be highly autonomous in this sense. By contrast, the relationship (call it (13)) between the pressure on the gas pedal and the speed of the car will be far less autonomous. We may imagine that (13) holds stably for some particular car if we intervene repeatedly to depress the pedal under sufficiently similar conditions. (13) will thus be invariant under some interventions. Nonetheless, (13) will be disrupted by all sort of changes—by variations in the incline along which the car travels, by changes in the head wind which the car faces, by changes in the fuel mixture that the car consumes, by changes in the internal structure of the car engine (e.g. by cleaning the spark plugs and adjusting the carburetor) and so on. (13) will also be disrupted by extreme interventions on the gas pedal—for example, those that are sufficiently forceful that they destroy the pedal mechanism. (13) is thus relatively fragile or non robust in the sense that it holds only in certain very specific background conditions and for a restricted range of interventions. Intuitively, although (13) is invariant under some interventions and changes, it is invariant under a 'smaller' set of interventions and changes than fundamental physical laws.

According to the account of explanation defended in Section **4** if (13) holds invariantly for some range of interventions that depress the gas pedal by

various amounts, for some type of car in a kind of environment, then we may appeal to (13) and to the depression of the pedal to explain the speed of the car, provided that the car is within the domain of invariance of (13). Within its domain of invariance (13) describes a relationship that can be exploited for purposes of manipulation and control—it describes how we can change the speed of the car by changing the depression of the gas pedal. This is a feature which (13) shares with paradigmatic laws like the gravitational inverse square law and which distinguishes both from purely accidental generalizations like (5) and (6). Because (13) is not completely lacking invariance, an explanation that appeals to (13) will exhibit, albeit in a very limited way, the pattern of active counterfactual dependence that I claimed in Section **4** was at the heart of successful explanation. We can appeal to (13) to explain even if, because of its relative fragility or for other reasons, we are unwilling to regard it as a law of nature. We can thus think of this example as illustrating my claim that it is invariance and not lawfulness *per se* that is crucial in explanation.

However, like Haavelmo I also take it to be obvious that an explanation of the speed of the car that appeals just to (13) is shallow and unilluminating. I follow Haavelmo in tracing this to the fact that the relation (13) is relatively fragile—it is invariant only over a very limited range of interventions and changes in background conditions and can be used to answer only a very limited range of what-if-things-had-been-different questions. A deeper explanation for the behavior of the car would need to appeal to laws and engineering principles (14)—like those mentioned by Haavelmo—that are invariant under a much wider range of changes and interventions. Not coincidentally such a deeper explanation is such that it could be used to answer a much wider range of what-if-things-has-been-different questions. For example, unlike (13) the generalizations (14) appealed to in this deeper explanation are such that they could be used to explain why the car moves with speed that it does over a variety of different kinds of terrain and road conditions, under a variety of different kinds of mechanical changes in the internal structure of the car and so on. The what-if-things-had-been-different account of explanation thus seems to capture the relevant features of Haavelmo's example in a very natural way.

What might it mean to say, as Haavelmo does, that one generalization, is invariant under a 'larger' set of changes or interventions than another? In Haavelmo's example, this question has a straightforward answer. To a very good degree of approximation, the range of changes and interventions over which (13) is invariant is a proper subset of the range of changes and interventions over which the generalizations (14) of the deeper engineering theory of the behavior of the car are invariant. That is, any change that will disrupt the latter will also disrupt (13) but not *vice versa*. Thus any properly behaved measure will assign a larger size to the domain of invariance of the latter.

A similar basis for comparison exists in the case of many other pairs of

generalizations. Compare the ideal gas law (4) with the van der Waals force law

$$[P + a/V^2][V - b] = RT \tag{15}$$

Here a and b are constants characteristic of each gas, with b depending on the diameter of the gas molecules and a on the long range attractive forces operating between them. For any given gas, the generalization (15) holds invariantly in circumstances in which the ideal gas law (4) holds, but it also holds invariantly in at least some circumstances—roughly those in which intermolecular attractive forces are important and in which the volume of the constituent molecules of gas are large in comparison with the volume of the gas—in which (4) breaks down. The range of changes or interventions over which (15) is invariant is again 'larger' than the range of changes over which (4) is invariant in the straightforward sense that the latter set of changes is a proper subset of the former. Moreover, just as in Haavelmo's example, this larger range of invariance means that we can use (15) to answer a larger set of what-if-things-had-been-different questions than (4). Thus we can use (15) to answer question not just about what would happen to the values of one of the variables P, V and T given changes in the others in circumstances in which intermolecular forces are unimportant and intermolecular distances large in comparison with molecular volumes, but also what would happen to P, V or T when these conditions no longer hold. We can also use the van der Waals equation to explain various phenomena having to do with phase transitions—again circumstances in which the simpler (4) breaks down. A similar relationship holds between many other pairs of generalizations—for example, between the laws of General Relativity and those of Newtonian gravitational theory.

It is important to understand that the claim that the range of changes and interventions over which a generalization (G_1) is invariant is a proper subset of the claim that a second generalization (G_2) is invariant is *not* merely a restatement of the claim that (G_1) and (G_2) are both true and that (G_1) is derivable from (G_2) but not *vice versa*. For one thing (G_1) may be derivable from (G_2) but not *vice versa* even if neither generalization is invariant at all. For example, the true generalization (G_1) that the all spatiotemporal regions of 1 meter radius within 10 light years of the earth contain cosmic background radiation at 2.7 degrees K is derivable from the generalization (G_2) that all spatiotemporal regions in the universe contain such background radiation but neither generalization is invariant—neither is change—relating and both apparently depend in an extremely sensitive way on the initial conditions obtaining in the early universe. As another illustration, Mendel's law of segregation is derivable from and not equivalent to the conjunction of Mendel's law and Galileo's law of freely falling bodies but the conjunctive

law is not invariant under a wider range of interventions than Mendel's law. When we compare generalizations with respect to range of invariance, we compare generalizations along a very specific dimension of generality. Such comparisons are very different from the comparisons that we make when we simply ask whether one generalization is derivable from another.

Although we may compare the range over which two generalizations are invariant when the proper subset relation just described holds, this obviously yields only a partial ordering. For many pairs of generalizations neither will have a range of invariance that is a proper subset of the other. Moreover, the proper subset relation provides at best a basis for ordinal comparisons. We can say that one generalization is invariant under a larger set of changes than another, but we have no basis for claiming that this set is large or 'important' (to use Haavelmo's word) in some more absolute sense. Is there some other basis on which we can make such claims? I believe that there is. This basic idea is more easily illustrated than precisely characterized, but the underlying intuition is this: for different sorts of generalizations, applicable to different sets of phenomena or subject matters, there often will be specific sorts of changes that are privileged or particularly important or significant from the point of view of the assessment of invariance—privileged in the sense that it is thought to be especially desirable to construct generalizations that are invariant under such changes and that generalizations that are invariant under such changes are regarded as having a fundamental explanatory status in comparison with generalizations that are not so invariant. The privileged changes in question will be subject matter or domain specific—one set of changes will be important in fundamental physics, another in evolutionary biology and yet another in microeconomics. Thus expectations about the sorts of changes over which fundamental relationships will be invariant help to set the explanatory agenda for different scientific disciplines. These expectations will in turn be grounded in very general empirical discoveries about the sorts of relationships in the domains of these disciplines have been found to be invariant in the past and under what sorts of changes.

As an illustration consider that in physics fundamental laws are expected to satisfy certain symmetry requirements. It is widely recognized that such symmetry requirements , especially when interpreted 'actively', are invariance requirements. They amount to the demand that fundamental laws remain invariant under certain kinds of changes—for example under spatial or temporal translation of a system of interest or under spatial rotations or under translation from one inertial frame to another (Lorentz–invariance). These demands are rooted in very general empirical facts about the natural world: that relationships can be found that are invariant under these changes and not others is an empirical discovery. These empirical discoveries in turn generate expectations about the kinds of symmetries physical laws should exhibit. At least at

present, generalizations that fail to satisfy such symmetry requirements are unlikely to be regarded as candidates for fundamental laws or explanatory principles, regardless of whatever other features they possess. The requirements thus have a special or privileged status—from the point of view of the assessment of invariance in physics they are more important than invariance under other sorts of changes. Unlike traditional accounts of laws, which leave it opaque why fundamental laws are expected to satisfy symmetry requirements, an invariance-based account makes these requirements intelligible.

For purposes of comparison, consider what counts an important kind of change for the purposes of assessing invariance in contemporary microeconomics. In microeconomics, individual economic agents are often assumed to conform to the behavioral generalizations comprising rational choice theory (RCT). For the purposes of this paper I will take these generalizations to include the principles of expected utility theory, as described, for example, in Luce and Raiffa ([1957]), together with the assumption that choices are self-interested in the sense that agents act so as to maximize some quantity which is directly related to their material interests, such as income or wealth. Even if we assume, for the sake of argument, that these generalizations are roughly accurate descriptions of the behavior of many participants in markets, it is clear that there are many changes and interventions over which these generalizations will fail to be invariant. For example, there are many pharmaceutical interventions and surgically produced changes in brain structure that will lead (and in some cases have led) previously selfish agents to act in non-self-interested ways or to violate such principles of RCT as preference transitivity. However, economists have not generally regarded these sorts of failures of invariance as interesting or important, at least if, as is often the case, they occur relatively rarely in the populations with what they deal.

By contrast, failures of invariance under other sorts of changes are regarded as much more important. For example, microeconomists often require that fundamental explanatory generalizations such as the principles of RCT be invariant under changes in information available to economic agents or under changes in their beliefs and under changes in the incentives or relative prices they face. Indeed, a standard assumption among many microeconomists—one might take it to be constitutive of a certain sort of methodological individualism—is that the generalizations that will be invariant under such changes in information and prices all describe the behavior of individual economic agents rather than the relations between macroeconomic or aggregate-level variables like ‘inflation’, ‘unemployment’, and ‘gross domestic product’. That is, the idea is that there are no purely macroeconomic relationships that are invariant under changes in information and incentives and hence that their are no fundamental explanatory relationships between macroeconomic variables.

As an illustration, consider the macroeconomic relationship known as the

Phillips curve. This describes the historically observed inverse relationship or trade-off between unemployment and inflation in many Western countries from the mid-nineteenth to mid-twentieth centuries. A crucial question is whether this relationship is (or was) invariant under policy interventions on these variables. According to some Keynesian models, the Philips curve does describe a relationship which is invariant under at least some governmental interventions that change the inflation rate. If so, governments would be able by increasing the inflation rate to decrease the unemployment rate—a highly desirable result. The burden of an influential criticism of these models developed by Lucas ([1983]) (the so-called Lucas critique) is that the relationship discovered by Philips is not invariant under such interventions—that the result of interventions that increase the inflation rate will not be to lower the unemployment rate but rather simply to produce changes in the Philips curve itself. Very roughly, according to this critique, increasing inflation will reduce unemployment only if employers or employees mistake an absolute increase in prices for a favorable shift in relative prices and (given the assumption that these agents are 'rational') this is not a mistake they will make systematically or for any length of time. As soon as these agents realize that a general increase in the price level has occurred or come to expect that such an increase will occur, unemployment will return to its original level. To put the point abstractly, the Philips curve is not invariant under changes in the information available to economic agents or under changes in their expectations of a sort that almost certainly will occur once the government begins to intervene to change the inflation rate. A similar point will hold for many other macroeconomic relationships.

This example illustrates how issues about invariance arise naturally in economics. The interesting question for economists is not whether the Philips curve is a law of nature or completely exceptionless but rather whether it is invariant under certain specific kinds of changes and interventions. If the Philips curve is not invariant under the relevant sorts of interventions, it will not be regarded as a fundamental economic relationship or as a relationship which it would be satisfactory to take as primitive in a deep economic explanation. (For example, if the Lucas critique is correct, it would be unsatisfactory to appeal to the inflation rate and the Philips curve to explain the unemployment rate.) This is not because it fails to be invariant under all possible changes and interventions (including all-out nuclear war or radical psycho-surgery performed on the entire US population) but because it (allegedly) fails to be invariant under a specific set of possible changes that are thought to be particularly important—changes in the information that economic agents receive.

My suggestion, then, is that both of the considerations described in this section—comparisons of invariance based on the proper subset relation and

judgments about the significance or importance of the intervention over which a generalization is invariant—play an important role in the construction and assessment of explanatory generalizations. Together they provide a basis for distinguishing among invariant generalizations with respect to degree and kind of invariance and for judging that although a generalization is invariant under some interventions, it is nonetheless relatively fragile or unrobust in the sense that it is stable only under an unimportant set of interventions or under a set of changes that is relatively small in comparison with some rival generalization. As remarked above, the idea that generalizations can differ from one another in the range or importance of the interventions over which they are invariant is one of a number of respects in which the invariance-based framework that I am recommending departs from the traditional framework for thinking about laws of nature and their role in explanation. In contrast to the traditional framework, which admits just two mutually exclusive possibilities (a generalization is either a law or else it is 'accidental'), the notion of invariance allows us to make a much richer set of distinctions among invariant generalizations. As we shall see, this makes the invariance-based framework much better suited for capturing the characteristics of explanatory generalizations in the special sciences.

6 Laws

In describing the difference between invariant and non-invariant generalizations and the differences among invariant generalizations, I have so far very largely avoided invoking the notion of a law of nature and the various standard criteria which philosophers have traditionally employed to distinguish between laws and accidental generalizations. These criteria take a variety of different forms: laws are said to be exceptionless generalizations representable by universally quantified conditionals, to contain only purely qualitative predicates and/or natural kind terms and to make no reference to particular objects or spatio-temporal locations, to have very wide scope, to support counterfactuals, to be projectable or confirmable by their instances, to be integrated or potentially integrable into a body of systematic theory and to play a unifying or systematizing role in inquiry. In order to have a useful shorthand way of referring to these criteria, let us call them the traditional criteria for nomological status.

What is the relationship between these criteria and the notion of invariance? My view, to be defended below, is that most of the criteria are not helpful either for understanding what is distinctive about laws of nature or for understanding the features that characterize explanatory generalizations in the special sciences. In general, it is the range of interventions and changes over which a generalization is invariant and not the traditional criteria that are crucial both to whether or not it is a law and to its explanatory status. Moreover, whether or

not a generalization is invariant (and if so, over what range of changes and interventions) is surprisingly independent of most of the traditional criteria—a point that is perhaps suggested by the fact that in appealing to the notion of invariance to describe the differences among various generalizations like (1), (2), (3), (4), (13), and (14) we did not need to explicitly invoke these traditional criteria at any point. In fact, a generalization may satisfy many of the traditional criteria and yet fail to be invariant and a generalization may be invariant even through it fails to meet many of the traditional criteria. Among the traditional criteria only one (support for counterfactuals) is relevant to whether a generalization counts as invariant, and even then, as we shall see below (Section **10**), this criterion is understood quite differently on the invariance-based approach than on the traditional approach.

One possible response to these facts is to drop the concept of a law of nature as unhelpful for understanding science and to focus directly on the notion of invariance since the latter notion captures, or so I have suggested, what is really relevant to successful explanation. This strategy should appeal to those philosophers (Cartwright [1983]; Giere [1988]; van Fraassen [1989]) who, on various grounds, have been skeptical of the whole idea that there are laws of nature or that the concept of natural law plays an important role in science. While I am by no means entirely unsympathetic to this suggestion, I will not adopt it in this paper. Instead, I will proceed on the assumption that the concept of a law of nature, although not an especially sharp or clear concept (see Section **11**), remains useful for understanding explanatory practice in some areas of science (principally physics and chemistry) although it is not very helpful in connection with in the special sciences. Continuing the precedent established in previous sections, I will follow standard scientific practice in describing paradigmatic generalizations from physics and chemistry like the gravitational inverse square law, Maxwell's equations, and the ideal gas law as genuine laws. However, I will argue that there is little motivation for extending the notion of law to cover all of the explanatory generalizations of the special sciences. Rather than thinking of all invariant generalizations as laws, I suggest instead that we think of laws as just one kind of invariant generalization. Laws do indeed play an important role in (some areas) of science, but they are both less central and less pervasive in science as a whole than traditional approaches suppose.

7 Invariance, qualitative predicates, and scope

I claimed above that whether a generalization is invariant is surprisingly independent of whether it satisfies most of traditional criteria for lawfulness. Showing this in detail would require a lengthy paper in its own right. In what follows I will illustrate this claim by focusing on just a few of these

criteria—the requirements that laws must not refer to particular places or times and must not be too narrow in scope (this section), the requirement that laws and explanatory generalizations must be exceptionless (Section **9**), and the idea that laws must support counterfactuals, which I discuss and reinterpret in Section **10**.

Consider first the common suggestion that laws must contain 'purely qualitative' predicates and must contain no essential reference to particular objects, times or places. As a number of writers have observed, this is a dubious criterion for lawfulness. It is thus of considerable interest that it is perfectly possible for a generalization to be invariant without satisfying this criterion. To see this, imagine that Clinton's pockets on January 8, 1999 do turn out to have the property that whenever a non-dime is introduced into them, it is turned into a dime. In such a case, the generalization (5) ('All the coins in Clinton's pockets on January 8 are dimes') would be invariant under such interventions despite the fact that it contains non-qualitative predicates and makes essential influence to a particular person and time. Indeed, we can imagine, consistently with the non-qualitative character of (5), that it is invariant under a very wide range of interventions and changes—that it continues to hold no matter how coins are introduced (and no matter which coins are introduced) into Clinton's pockets, and no matter which other background changes occur.

This particular example is of course fantastic but there are others that are less so. Aristotle is often represented as thinking that as a matter of law (16) all freely falling objects will move toward particular spatial location—the center of the earth. He was, of course, wrong about this, but what was the nature of his mistake? If it is built into the very concept of law of nature that laws must not refer to particular places, his mistake was a conceptual one. A much more plausible judgment is that his mistake was empirical. The connection between lawfulness and invariance for which I have been arguing supports this judgment, for it is clear that (16) might have turned out to be a highly invariant generalization despite its reference to a particular object or spatial location.

For similar reasons, it is perfectly possible for a generalization to be invariant only for changes and interventions that occur within a limited spatial or temporal interval and to break down outside that interval. Suppose that, contrary to actual fact, the Phillips curve turned out to be invariant under governmental interventions that changed the inflation rate between, say, 1870 and 1970 in the United *S*tates, although not invariant outside this interval. If this had been the case, then (I would claim) despite the limited spatio-temporal scope of this relationship, one could appeal to it and to the fact that the US government intervened to raised the inflation rate in 1915 to explain why unemployment fell after this intervention. More generally, in contrast to traditional law-based accounts of explanation, the notion of invariance allows us to talk about explanatory relations that hold only over limited

spatio-temporal intervals or which make reference to particular objects, events or processes. As we shall see below, many explanatory generalizations in the special sciences seem to have exactly these features and this is one reason why the notion of invariance is particularly well suited to understanding their character.

Consider next the relationship between invariance and scope. This latter notion is difficult to characterize precisely, but I will take the intuitive idea to be that a generalization has wide scope if it holds for a 'large' range of different kinds of systems, in the sense that the systems in question satisfy both its antecedent and its consequent and that we can appeal to the generalization to explain the behavior of such systems. For example, we think of the Newtonian inverse square law as having wide scope because it holds for all masses throughout the universe—for bodies falling near the surface of the earth, for all planets orbiting the sun and so on. By contrast, we think of a version of Hooke's law (17) $F=-K_sx$ that describes the behavior of one particular sort of spring, characterized by the specific spring constant K_s, as much narrower in scope. Most other sorts of springs will obey a different (or no) version of Hooke's law, and most systems that are not springs will not be describable in terms of this law at all.

While I think that it is true (see Section **11**) that we are reluctant to regard generalizations with narrow scope as laws, it is nonetheless of considerable interest that the scope of a generalization seems to have little to do with whether it is invariant and, if so, over what range of changes and hence on my view, little to do with explanatory import.[5] Despite its narrowness of scope, the generalization (17) might well turn out to be invariant under a substantial

[5] The notion of scope and its relation to invariance deserve a more detailed treatment than I can give them here. Intuitively, the scope of (17) has to do with how many different kinds of springs (or perhaps, alternatively, with how many individual springs) are correctly described by (17). For example, the scope of (17) would be greater if it correctly described not just springs made out of the material that is characteristic of S but also other sorts of springs, found elsewhere in the universe and made of very different kinds of material. The scope of (17) would be narrow if there were only one spring in the universe conforming to (17) and greater if such springs were very common. By contrast, when we ask about the range of invariance of (17) *qua* description of the behavior of S we are asking a different kind of question. In this case, we want to know, for springs of this very type (regardless of how many or how many different kinds there are), the range of interventions over which (17) is invariant. Conversely, even if there was only one spring in the universe conforming to (17), (17) could still be a highly invariant generalization concerning the behavior of that spring.

In general, scope differs from invariance in at least two ways. First, invariance is a *modal* notion—it has to do with whether a relationship would remain stable under various hypothetical changes. In contrast, scope is understood in actualist, non-modal terms—it has to do with how many systems or how many different kinds of systems there actually are to which a generalization applies. Second, invariance requires stability under interventions—where there is no well-defined notion of intervention or change there is no notion of invariance to apply. Suppose a generalization like (17) describes the behavior of springs made of two different kinds of materials M1(plastic) and M2 (copper). There may be no well-defined notion of changing a spring made of M1 into one made of M2 or at least no way of carrying out the change so that (17) is stable over intermediate steps in the change. (As Hitchcock and Woodward [unpublished manuscript] put it, (17) isn't stable in a neighborhood around M1.) Hence one can't legitimately talk about (17)

range of interventions that change the extension of a particular spring or set of springs and under other changes in background circumstances as well. If so, according to the account of explanation advocated here, we can appeal to this generalization to explain why a particular spring exerts the force that it does. Again this is important for understanding explanation in the special sciences. Many generalizations in the special sciences, such as the regression equations described in Section **12** or generalizations about particular biological mechanisms lack broad scope—intuitively, they are about very specialized kinds of systems—but it would be a mistake to conclude on this ground alone that they are unexplanatory.

8 Invariance and exceptionlessness

I turn next to a more extended discussion of one of the most important of the traditional criteria for lawfulness. This is the requirement that laws must be exceptionless.[6] Most philosophers still endorse this idea; a particularly straightforward expression can be found in a recent paper by Paul Pietroski and Georges Rey:

> The key feature [of laws] [. . .] is the universal quantifier. Laws say that *whenever* some initial condition obtains, some other condition obtains as well. A single instance of [F.-G] shows the generalization 'F→G' to be false, in which case *a fortiori* there is no such law ([1995], p. 83).

If taken literally, the requirement that genuine laws must be exceptionless virtually forces the law/accident dichotomy on us since exceptionlessness is an all or nothing matter, and not one of degree. This is also the requirement that creates some of the deepest difficulties for the contention that explanatory generalizations in the special sciences are laws since, on the face of things, most such generalizations seem far from exceptionless. The centrality of this

being invariant under changes from M1 to M2. If (17) applies to both M1 and M2, it has broader scope than if it applies just to springs made of M1, but it isn't any more invariant.

One of the many reasons why this distinction is important is its bearing on unificationist views of explanation like those defended by Friedman ([1974]) and Kitcher ([1989]). In my view it is counterintuitive that the goodness of the explanation of the behavior of some particular spring should depend on the scope of (17). One does not understand the behavior of the spring better if it should turn out that there are many other systems or kinds of systems to which (17) applies rather than just one. However, for all of the reasons adduced above, it is not at all counterintuitive that the goodness of this explanation has something to do with the range of changes and interventions over which (17) is invariant. From my perspective, the unificationist account of explanation, at least as formulated by Friedman and Kitcher, fails to distinguish between scope and range of invariance; it mistakenly takes the degree of unification a generalization achieves and hence its explanatory import to depend on its scope.

[6] Systems that represent exceptions to a generalization most of course be distinguished from systems to which the generalization merely fails to apply. I will follow Pietroski and Rey in thinking of exceptions as involving cases in which the behavior of a system satisfies the antecedent of a generalization but not its consequent. By contrast, a generalization will fail to apply to a system if it fails to satisfy the conditions specified in its antecedent.

requirement in contemporary discussions of laws and explanation in the special sciences is indicated by the existence of a very substantial literature on ceteris paribus laws, to be examined in more detail in Section **13**, that is premised on the assumption that to vindicate the lawfulness and explanatory status of generalizations in the special sciences one must show that these generalizations are (or can be closely associated with 'backing' generalizations that are) exceptionless. Needless to say, there would be no motivation at all for this project if (as I shall argue) it is a mistake to suppose that to qualify as a law or as invariant or explanatory a generalization must be exceptionless.

By way of contrast with the traditional view that exceptionlessness is essential for lawfulness, I said, in describing the ideal gas law (4), that it was invariant under a certain range of changes in the variables *P*, *V*, and *T* but broke down or failed to hold exceptionlessly under others (for example, under conditions like extremely high pressures at which intermolecular forces become important). I thus took it that (4) could count as a genuine law and figure in explanations of the behavior of gases within the domain over which it is invariant even though it had exceptions outside of this domain. Philosophers, who like Pietroski and Rey, require that genuine laws are exceptionless must hold that (4), as it stands, is no law. The usual move is to suggest that, insofar as there is a law associated with (4), this will be a generalization that incorporates some appropriate set of qualifications and conditions into its antecedent in such a way as to render it exceptionless. Thus it will be suggested that the law associated with (4), is not really expressed by (4) itself, but rather by some more complicated, exceptionless generalization like (18) 'In circumstances $C_1 \ldots C_n$ (where $C_1 \ldots C_n$ are taken to exclude the possibility that intermolecular forces are important etc.), (4) holds'.

In fact, however, most known examples of physical laws follow a pattern like the one I have attributed to the ideal gas law. They are invariant only under a certain domain or regime of changes and break down outside of these. For example, the laws of classical electromagnetism (Maxwell's equations) break down at scales at which quantum mechanical effects become important. Similarly, the field equation of general relativity are widely expected to break down at very small distances (the so-called Planck length) at which quantum gravitational effects become important.

In my view, we should resist the conclusion that these facts show that Maxwell's equations and the field equations of general relativity are not, in their usual formulations, genuine laws and that the genuine laws associated with these generalizations are instead exceptionless generalizations constructed on the model of (18) (i.e. generalizations like (19) 'If such and such conditions are satisfied, then Maxwell's equations will hold'). To begin with, such a view is sharply at odds with standard scientific practice which is to take Maxwell's equations and the field equations as laws just as they stand and to

appeal to these generalizations, rather than exceptionless reconstructions of them like (18)–(19), in order to explain. Moreover, scientists often make use of laws in their usual form (that is, with exceptions) in circumstances in which they are unable to describe in a precise or theoretically perspicuous way, the exact boundaries of the domains over which they are invariant—that is, in circumstances in which they do not know how to turn them into an exceptionless generalizations along the lines of (18)–(19). For example, scientists regarded Maxwell's equations as laws of nature and appealed to them to explain long before they were able to correctly specify the circumstances in which these equations break down. It is a reasonable guess that many generalizations presently regarded as laws similarly will be found to break down in circumstances that are not at present understood. Indeed, some theorists claim that this will be true for all laws of nature—a suggestion that is incoherent on the traditional account of laws, although not on the invariance-based account. Even if we put this possibility aside, it seems clear that if we demand that all genuine laws must be exceptionless generalizations, it follows that we know very few laws and, given the additional assumption that explanations must cite laws, that most of the generalizations we know how to formulate, even in physics, cannot be used to explain. Finally, and most importantly, we shall see in the following section that there are principled reasons, grounded in the kinds of information that we expect successful explanations to provide, why generalizations like the ideal gas laws or Maxwell's equations are formulated in their usual, exceptioned form, rather than in the exceptionless form (18)–(19).

I believe that it is an important advantage of the notion of invariance that it provides a way of capturing this feature of laws—that it doesn't require that laws be exceptionless. As (4) illustrates, it makes perfectly good sense to think of a generalization as invariant across a certain range of changes and interventions even if it is not exceptionless or invariant across all such changes. The idea that Maxwell's equations are invariant across certain kinds of changes or conditions but not others and that this sort of invariance is sufficient for those equations to count as laws and to figure in explanations represents in a natural way the role that these generalizations actually play in scientific practice.

9 Exception incorporation *vs.* independent specification

I suggested above that there are principled reasons, in addition to the considerations rehearsed in the previous section, why it is appropriate to think of generalizations like Maxwell's equations or the field equations of general relativity as laws just as they stand and misguided to demand that all laws or explanatory generalizations must be exceptionless. It will be useful to explore these in more detail since they will play an important role in our subsequent discussion.

Let me begin by raising what might seem to be an obvious objection to the position I have been defending. Consider the formulation that I favor, in which we claim that (20) some generalization (G) is invariant within a certain domain D but breaks down or has exceptions outside this domain. I will call this the independent specification model since the domain D in which (G) holds is specified independently of (G), rather than being packed into the antecedent of (G). The objection to this goes as follows: given the information embodied in (20), can't we always reformulate this model along the lines suggested in the previous section, as an exceptionless generalization of the traditional kind, the antecedent of which is restricted to D? That is, why not think of (20) as equivalent to an exceptionless generalization which says (roughly) (21) 'For all *X*, whenever *X* is in domain D (or satisfies whatever conditions are sufficient for being in domain D), then [. . .] (here follows the original generalization G)'. I will call this the exception-incorporating model, since the restrictions on D are incorporated directly into the generalization (21) rather than being specified independently.

Even if it is true that generalizations in science are typically formulated along the lines of the independent specification model (20), rather then along the lines of the exception-incorporating (21), it is tempting to think of (20) and (21) as two different ways of representing the same claim—as mere notational variants on one another. Isn't it arbitrary whether we specify the domain independently of (G), as (20) does, or build it into the antecedent of a generalization as (21) does? If anything, isn't (21) simpler and more perspicuous than (20) and more in accord with traditional ideas about the features laws must possess?

In what follows I will argue that the formulations (20) and (21) are not equivalent or interchangeable: the two formulations are motivated by quite different views about the sort of information that is required to improve an explanation and about what one needs to know in order to successfully explain. They are also associated with two quite different ways of thinking about the content of scientific theories. There are in fact good reasons for preferring (20) to (21).

Let me begin with an observation designed to undermine the suggestion that the exception-incorporating formulation (21) is automatically more natural or perspicuous. Once one gives up the idea that successful explanation is just a matter of nomic subsumption (or of showing that an explanandum was to be expected), it will not be always or automatically true that we improve the explanatory credentials of a generalization with exceptions by replacing it with a generalization that is exceptionless or more nearly exceptionless. In particular, according to the account of explanation defended in Section **4**, changes to a generalization that render it exceptionless (or more nearly exceptionless) but do not enable it to figure in the answers to a larger range of what-if-things-

had-been-different questions will not constitute an explanatory improvement. It is in part because of this that exceptionlessness is a less crucial feature of laws and explanatory generalizations than many philosophers have supposed.

As an illustration, suppose that, as we have imagined, General Relativity does break down below the Planck length but only in those circumstances, so that the generalization (22) 'Above the Planck length (here follow the field equations)' is genuinely exceptionless. Consider some phenomenon such as the deflection of starlight by the sun which we ordinarily take to be explained by General Relativity in its usual, exceptioned form—i.e. just by the field equations themselves. Would this explanation be improved if we were to replace the field equations with the exceptionless generalization (18)? Not (or at least not obviously) according to the account of explanation sketched in Section **4**. The reason is that an explanation of starlight deflection in terms of (22) does not convey new information in addition to what is conveyed by the field equations themselves about what would happen if the conditions cited in its explanans were to change or at least it does not convey the kind of precise or specific information about this that we desire in a successful explanation. The explanation in terms of (22) tells us nothing definite about what would happen if the additional condition added to (22)—being above the Planck length—were to change. That is, we are told nothing specific about what would happen under conditions below the Planck Length, other than (by implication) that the field equations will no longer hold.

The idea that (22) does not represent a serious explanatory advance on the field equations as ordinarily formulated may seem very odd to those whose judgments about such issues have been nurtured by nomic subsumption models of explanation and by the closely associated idea that laws must be exceptionless. None the less, I submit that this judgment is just scientific common sense. A genuine explanatory advance over General Relativity would require the actual construction of a unified theory of gravity which embraces both quantum and macroscopic gravitational phenomena. Presumably such a theory would show in terms of some single set of principles, both how gravitational phenomena behave at very small length scales and how General Relativity turns out to be correct or nearly correct at large distances. Such a theory could thus be used to answer a larger set of (interesting) what-if-things-had-been-different questions than General Relativity and for this reason would represent an explanatory advance. However, merely to specify, as (22) does, that GR breaks down below the Planck Length is not to provide or exhibit such a unified theory (although it no doubt suggests something about the form such a theory will take). It is in part because the substitution of (22) for the field equations represents no serious explanatory advance that scientists are usually quite happy to employ the usual formulation of GR rather than (22) to explain phenomena that fall within the domain of GR.

To put the point in a way that anticipates my discussion below, the condition 'being above the Planck Length' plays a different role than the causal or explanatory factors that figure in the field equations such as the stress-energy tensor or the curvature tensor. In contrast to the mass-energy distribution in some region of spacetime which genuinely helps to explain why this region has a certain curvature, 'being above the Planck Length' doesn't describe a factor that explains anything or makes anything happen. Its role is rather to describe a condition for the application of GR or to help specify the domain over which GR holds. Scientists recognize this different role by not thinking of this condition as built into the field equations themselves but rather as specified independently, along the lines of (20) above.

There are other considerations, also rooted in scientific practice, that support this analysis. The idea that laws and other explanatory generalizations must be exceptionless goes along with (is supported by) a certain picture of how theorizing and model building work in science: a picture according to which explanatory generalizations already contain (at least if properly formulated) within themselves full specifications of their exact domains of application. However natural this picture may seem to philosophers, there is little doubt that it is a descriptively inaccurate account of scientific practice. A more descriptively realistic picture is this: at any given time, scientists will have in their possession various generalizations which they have successfully used to model and explain certain phenomena. However, it is typically a separate empirical question, the answer to which is not already built in to these generalizations, what the full range of phenomena is that can be so explained. Sometimes, as in the case of GR, scientists can give a simple, precise, and general characterization of the domain in which a generalization holds. Then one can read off just from this characterization whether the generalization can be appropriately applied to some potential explanandum. But more commonly, especially in the special sciences, such a general characterization will be unknown and may not even exist at the level at which one is theorizing. Instead, the circumstances in which generalization breaks down will be very complex and heterogeneous and in many cases not known with any precision. In this sort of case, whether the generalization holds for various previously unexplained phenomena must be discovered empirically, on a case by case basis, by seeing whether the generalization can be successfully applied to them, perhaps with the guidance of some very rough rules of thumb.

As an illustration of this point, consider Mendel's law of segregation which is standardly formulated as the claim that in sexually reproducing organisms each gene from a pair has 0.5 probability of being represented in a gamete. When so formulated this generalization breaks down in a number of different circumstances—for example, when meiotic drive is present. However, the law of segregation does not by itself tell us what these circumstances are—instead

whether substantial violations are present in particular populations often must be determined 'empirically' on a case by case basis. L. C. Dunn, the discoverer that the t-allele in house mice (which is responsible for taillessness) does not conform to Mendelian segregation, describes this feature of biological practice when he writes:

> Mendelian heredity and its corollary, Hardy–Weinberg equilibrium in panmitic populations, assume [that the probabilities of the A and a gametes produced by the heterozygote Aa are equal] as a matter of course and the assumption is generally justified by direct evidence and by success in application. But the rule is not universal [. . .] (Dunn [1957], pp. 139–40, quoted in Beatty [1979], pp. 131–2).

In contrast to the exception-incorporating model, Dunn does not think that the law of segregation is exceptionless or that one can determine whether some particular trait conforms to the law merely by examining whether the conditions specified in the antecedent of the law are satisfied. Instead such a determination is made on the basis of additional evidence or 'success in application'.

As a second illustration consider the principles of rational choice theory. In addition to the violations of these principles discussed in Section **5** there is general agreement that exceptions are more extensive in connection with certain kinds of political and economic phenomena than others. For example, in the course of a recent defense of rational choice models, Fiorina ([1995], p. 88) claims that:

> RC [Rational Choice] models are most useful where stakes are high and numbers low, in recognition that it is not rational to go to the trouble to maximize if the consequences are trivial and/or your actions make no difference.

In a similar vein, Green and Shapiro ([1995], p. 267) write, in the course of a critical survey of RCT:

> Rational choice explanations should be expected, prima facie, to perform well to the extent that the following five conditions are met: (i) the stakes are high and the players are self-conscious optimizers; (ii) preferences are well ordered and relatively fixed (which in turn may require actors to be individuals or homogeneous corporate agents); (iii) actors are presented with a clear range of options and little opportunity for strategic innovation, (iv) the strategic complexity of the situation is not overwhelmingly great for the actors, nor are there significant differences in their strategic capacities, and (v) the actors have the capacity to learn from feedback in the environment and adapt. Our conjecture is at bottom empirical, rooted in our best judgment concerning why rational choice models have failed in the literatures we have examined. We might be wrong about one or more of these constraints; only the progress of empirical inquiry will tell.

For example, rational choice models typically provide better explanations and more accurate predictions of the behavior of political *élites* and party leaders (who are often in a position to expect a strong influence on outcomes about which they are informed and care a great deal) than of the decisions of individual voters, who are often not well-informed and whose chances of casting a decisive ballot are typically extremely small. For similar reasons, as Satz and Ferejohn observe in a recent paper ([1994]), rational choice models have been far more successful in explaining the behavior of firms than the behavior of individual consumers.

In the passages quoted above, the likely domain of RCT and the circumstances in which it is likely to break down, are specified in an informal and rather imprecise way, and independently of the basic explanatory principles of RCT rather than being incorporated into those principles. That is, they follow the pattern of the independent specification model (20) above, rather than the exception-incorporating model (21). One reason why it is implausible to suppose that, appearances to the contrary, these restrictions should be regarded as built into the principles of RCT is that the restrictions are inconsistent with the principles if the latter are interpreted as universal laws. For example, RC principles themselves tell us that we should not expect that people's behavior will be any less self-interested when stakes are low than when they are high. It is precisely because this is the case that explaining why most people bother to vote creates such difficulties for RC approaches. In addition, the restrictions described above are obviously vague and imprecise—they are best viewed as rules of thumb rather then as specifications of the exact circumstances in which we should expect RC principles to hold. This imprecision makes the restrictions unattractive candidates for incorporation into the antecedents of RC principles themselves—a point to which I will return below. Finally, as the above quotations make clear, the restrictions represent empirical discoveries that result from a long series of unsuccessful attempts to apply rational choice approaches to the phenomena described above—attempts that clearly reveal the extent to which those who use the theory do not regard a specification of the phenomena for which the theory holds as built into the fundamental principles of the theory.

While model (20) in which domain and generalization are specified independently often seems to provide a more accurate description of scientific practice, one might still wonder whether the exception-incorporating model (21) is more normatively perspicuous. In what follows, I will argue that the reason that scientific practice conforms to model (20) is that this model also has certain normative advantages: it allows us to formulate a much more plausible account of what one needs to know in order successfully to explain. Implicit in the account of explanation defended above are the following epistemic requirements: if one wishes to explain the behavior of system S, one needs to know

(23) an invariant generalization (G) and (24) information about initial conditions holding in S that when combined with (G) can be used to answer a range of what-if-things-had-been-different questions about the behavior of S. Knowing this requires knowing (25) that with respect to the behavior in question S is in the domain of invariance of (G). However, in order to use (G) to explain one need not know (26) the exact boundaries of the domain D over which (G) holds—i.e. one doesn't have to know an exceptionless version of (G).

This conception fits naturally with the independent specification model (20) and with the remarks from Fiorina and Green and Shapiro quoted above. The idea is that one can appeal to the principles of RCT to explain the behavior of, say, buyers and sellers in a certain market, as long as one knows that the behavior of these agents is within the domain of invariance of these principles (that is, as long as these principles correctly describe how those participants would behave under some relevant range of changes in variables like prices) even though one is unable to state these principles in a completely exceptionless form and they almost certainly break down in unknown ways for some other actors in other circumstances. That is, the fact that such principles are violated by, say, ordinary voting behavior doesn't undercut their use to explain behavior in domains in which they are invariant. Similarly, a Nineteenth Century physicist can use Maxwell's equations to explain various classical electromagnetic phenomena while having false beliefs about (or no beliefs at all about) the conditions under which Maxwell's equations fail to hold—while being unable to formulate Maxwell's equations in a genuinely exceptionless way—as long as it is known to be true that these classical phenomena fall within the domain of invariance of Maxwell's theory. The independent specification model permits us to distinguish between, on one hand, knowing (23)–(25) and, on the other hand, knowing (26) because we don't think of the information (26) as already built into (G) and because we can know that (25) some system of interest is within the domain of (G) without knowing the exact boundaries of that domain. It thus allows us to express the idea that it is invariance rather than exceptionlessness that is crucial to successful explanation. By contrast, on the exception-incorporating model, no such distinction is available. Information about the boundaries of D must be built into (G) itself and if, as is typically the case, one doesn't have such information, one will typically be unable to formulate (G) itself in an acceptable way.

While the exception-incorporating model (21) connects successful explanation with the possession of exceptionless generalizations, the independent specification model (20) fits more naturally with the undeniable fact that, especially in the special sciences, in constructing explanations we often must appeal to generalizations the exact boundaries of whose domains are unknown or very difficult to characterize in a precise way. My suggestion is

that it is part of our methodology for constructing and evaluating explanations that this sort of imprecision is allowable in the specification of the domain over which an explanatory generalization holds but not acceptable when specifying the generalization itself. The informal qualitative descriptions of the domain over which rational choice theory may be expected to hold quoted above illustrate this basic idea—the imprecision of such descriptions is acceptable when characterizing the domain of the theory, but would be unacceptable if built into the fundamental generalizations of RCT. When our knowledge of the limits of validity of a generalization are vague, or when we know or suspect it doesn't hold exceptionlessly but are unable fully to enumerate the exceptions, we build the vagueness into our characterization of its domain rather than building it into the antecedent of the generalization itself. Nor is this arbitrary: as suggested above we can operate perfectly well with domains that have vague boundaries since we can often know that we are within those boundaries even if we don't know exactly what they are. By contrast when vague and unclear domain restrictions are built into the antecedent of a generalization we are left with a candidate for a law without any definite content at all.

10 Invariance and counterfactuals

The notion of invariance is obviously a modal or counterfactual notion—it has to do with whether a relationship would remain stable if, perhaps contrary to actual fact, certain changes or interventions were to occur. What is the relationship between the idea that laws (and other explanatory generalizations) describe invariant relationships and the more familiar idea, defended by many writers, that what distinguishes laws from accidental generalization is that the former but not the latter 'support' counterfactuals? Does the invocation of invariance merely restate this more familiar idea in slightly different language? I claim that the answer to this question is 'no'. The relationship between laws (and other explanatory generalizations) and counterfactuals suggested by the notion of invariance is interestingly different from the standard philosophical picture of their relationship.

Consider a generalization of form

(27) All *A*s are *B*s

and the associated counterfactual

(28) If *X* were to be an *A*, then it would be a *B*.

The standard philosophical view is that if (27) is a law, it will support counterfactuals of form (28); by contrast if (27) is accidental, it will fail to support such counterfactuals. Exactly what 'support' means and exactly which

counterfactuals of form (28) must be supported is typically left unclear; in practice the requirement is often interpreted in such a way that (27) counts as a law if we can find some true counterfactual (or perhaps some small set of true counterfactuals) to associate with it.[7]

There is an obvious problem with this proposal. There seem to be many generalizations that in some sense support[8] counterfactuals of form (28), but that are plainly not laws, or even invariant generalizations. Consider a variant of an example due to Aardon Lyon ([1977]). A museum has adopted a policy such that

(29) All of the Sisleys in its possession are hung in room 18.

You are ignorant of this policy and ask, regarding some painting in room 17, whether it is a Sisley. You are told in response:

(30) If this painting were a Sisley, then it would be in room 18.

There is a natural reading of the counterfactual (30) according to which it is true and according to which it is supported by the generalization (29). Similarly, suppose for the sake of argument that

(31) All drivers in England drive on the left hand side of the road.

There is a natural reading of the counterfactual

(32) If I were to drive in England, then I would drive on the left hand side of the road

according to which (32) is true and according to which it is supported by (31). Indeed, it is easy to imagine circumstances in when even the generalization

(5) All the coins in Clinton's pocket are dimes.

'supports' (or appears to support) some counterfactuals of form (28). Suppose (5) is found to be true over an extended period of time and in circumstances in which many coins move in and out of Clinton's pocket. Then it is a very reasonable inference that there is some systematic reason or cause why (5) is true. Perhaps, for example, Clinton has made it his policy to allow only dimes in his pocket. Given that this is the case, if you are told that some particular coin c is in Clinton's pocket, you have good reason to suppose that it is a dime. You thus have good reason to accept the counterfactual

(33) If c were a coin in Clinton's pocket, then it would be a dime.

[7] For example, this is the *de facto* position adopted in Fodor ([1991]) and Pietroski and Rey ([1995]).

[8] One possible response to the examples that follow is to try to develop a more precise notion of support according to which the counterfactuals described below are not really supported by the generalizations associated with them. I shall not explore this alternative. My suspicion is that, if developed adequately, it will coincide with the position I defend.

provided this is understood in a what we earlier called a passive sense (that is, as a claim about what it would be reasonable to believe about c, given the information that it is in Clinton's pocket) rather than in an active sense (as a claim about what would happen to a non-dime if it were introduced into Clinton's pocket as a result of intervention).

Nonetheless, (29), (31), and (5) are plainly not laws of nature. Indeed, both (29) and (5), at least, strike us as paradigmatic examples of accidental, non-explanatory generalizations. A focus on invariance explains the basis for these judgments. Even if (5), (29), and (31) are true, our background knowledge leads us to believe that all three generalizations are highly non-invariant in the sense that they would breakdown under many different sorts of interventions and changes in background conditions. For example, (29) would break down under interventions that consist in introducing a Sisley into room 17 or under other changes in the museum's policy regarding the hanging of pictures. (31) would no longer hold if the British government were to decide (perhaps in order to bring British driving practices into conformity with those in the US and continental Europe) that henceforth British motorists must drive on the right. In addition individual motorists might, either as a result of inattention or deliberate decision, drive on the right, thus disrupting (31). Finally, as already noted, (5) will break down under interventions that consist of the introduction of a non-dime coin into Clinton's pocket.

Broadly speaking, examples like those just described seem to work in the following way. We have a generalization that holds because certain background conditions or generating structures are in place (for example, a decision to hang all Sisleys in room 18 or the existence of a convention that everyone drives on the left, to which each driver finds it rational to conform, given that others conform). Such generalizations are fragile or non-robust in the sense if these background or generating conditions were to change the generalizations would no longer hold—hence they are not invariant under such changes. Nonetheless we may have very good reasons for thinking these background conditions will in fact persist—(e.g. that the left hand side convention will remain in force) and hence that the generalizations in question will continue to hold. In these circumstances it will often be very plausible to accept counterfactuals like (30) and (32) and (33) (at least when interpreted passively) that are associated with these generalizations despite their non-invariance.

These examples illustrate how the traditional counterfactual test for lawfulness embodied in the relationship between (27) and (28) and the test associated with invariance come apart—a generalization can pass the former test but not the latter. The requirement that laws and other explanatory generalizations be invariant differs from the traditional idea that what is distinctive about laws is that they will support counterfactuals of form (28) in at least two respects. First, on the invariance-based approach we attach a

special significance to a particular sort of counterfactual—active counterfactuals whose antecedents describe interventions that realize or bring about *A*. As the above examples (as well as our discussion in Section **3**) illustrate, a generalization can support passive counterfactuals without supporting any active counterfactuals (and hence without being invariant under any interventions) at all. For example, although (29) supports some counterfactuals, it fails to be invariant under interventions because it fails to support active counterfactuals like

(34) If one were to introduce a Sisley into the museum *via* an intervention, then it would be in room 18.

This illustrates one respect in which the counterfactual test associated with the notion of invariance is stronger then the traditional requirement that a generalization of form (27) support some counterfactuals of form (28).

A second difference is this: on the traditional approach, one considers just a single counterfactual of form (28) and asks, whether such a counterfactual is supported by (27)—a question which presumably requires a single yes or no answer.[9] (This reflects the dichotomous character of the traditional law/accident framework.) One assumes that a generalization either passes the test of supporting (some) counterfactuals, and hence qualifies as a law, or else it fails this test and is accidental. By contrast, as I have emphasized, a generalization can be more or less invariant and it can be invariant under one set of interventions or changes and not under others. Instead of associating a kind of single counterfactual with a generalization like (27), the invariance-based approach suggests instead that we think in terms of a whole family of counterfactuals whose antecedents correspond to different interventions or different changes in background circumstances under which *A* would be true and then ask whether *B* would also be true in those circumstances. Some of these counterfactuals may turn out to be true and others false and together they give us the range of circumstances or interventions over which the generalization is invariant. We thus do not confine our attention to a single world or set of worlds that is 'closest' (however this is defined) to the actual world in which *A* is true and ask which *B* is true in that world but rather regard it as legitimate to consider counterfactuals in which *A* occurs under conditions that are very dissimilar from those that hold in the actual world. For example, in considering the range of changes over which the field equations of General Relativity are invariant, it is perfectly appropriate to consider worlds in which the mass distribution of the universe is extremely different from that obtaining in the actual world (for example, worlds consisting of a single star or which are

[9] Alternatively, if one considers a set of counterfactuals the relevant counterfactual test is usually taken to be whether (27) will support all counterfactuals in the set—again a question which must receive a single yes or no answer.

entirely empty of mass) and to ask whether the equations of GR would continue to hold under such circumstances. Indeed, it is standard practice to assume that those equations are highly invariant in the sense that they would continue to hold under such extreme circumstances.

Similarly, although the automotive generalization (13) relating gas pedal position and speed in a particular make of car may hold for other cars of that make in circumstances that are close to the actual circumstances, so that in this sense (13) is counterfactual-supporting, in assessing the range of invariance of (13) we also consider what would happen under cases in which the internal structure of the car and other features of its environment are changed in more radical ways. It is the fact that we think that (13) would break down under many such counterfactual circumstances that distinguishes it from less fragile generalizations like the field equations. Similarly, in evaluating whether a generalization like (31) is invariant, we consider not just whether (31) would continue to hold in a world which is as much like the actual world as possible, consistent with my driving in England, but also consider whether (31) would continue to hold in circumstances that depart much more substantially from the actual world. Thus we consider worlds in which, e.g. the British government issues a decree requiring that everyone drive on the right, or in which Britain joins the United States and adopts American driving habits or in which I drive in England but no longer care about my own safety or that of others. To put the point in a slightly different way, when we focus on invariance we replace a single counterfactual of form (28) with a whole family or series of counterfactuals corresponding to different ways of strengthening the antecedent of (28)—we consider a whole set of counterfactuals of the form

(28*) If *A* and *C* were true, then *B* would still be true.

where different possible interventions bring about *A* and *C* represents different possible background circumstances in which *A* might occur.

11 Are all invariant generalizations laws?

I suggested above that there are invariant generalizations that are not naturally regarded as laws, if one has even a modestly demanding conception of what a law is. Section **13** will discuss some examples in more detail. Nevertheless, it will be useful at this point to introduce a simple physical example that illustrates the contrast I have in mind. My object in doing so is not to legislate regarding the proper use of the word 'law' but rather to draw attention to some differences between paradigmatic laws and other sorts of invariant generalizations.

Consider again a particular sort of spring *S* which, over a certain range of extensions, conforms to Hooke's law (17) $F=-K_sX$, where X is the extension of

S, *F* the restoring force it exerts, and K_s a constant characterizing *S*. Suppose that (17) is invariant under some interventions that change the extension of *S*, but breaks down for extensions that are too large and for other sorts of changes in background conditions as well. As we have seen, this fact by itself does not distinguish (17) from paradigmatic laws of nature, such as Maxwell's equations. Nonetheless, there appear to be several respects in which (17) does differ from paradigmatic laws. First, as already discussed, (17) is much narrower in scope. A second difference, which is perhaps more significant from the point of view of explanation is this: not only will there be a range of 'extreme' extensions for which (17) breaks down but even if we confine our attention to extensions of *S* that are not in this range—i.e. extensions for which (17) sometimes holds—there will be a large number of possible changes in background conditions that do not explicitly figure in (17), for which (17) will be violated. For example, even if it is true that in 'normal' circumstances *S* will conform invariantly to (17) under small extensions, it will not do so if it is heated to a high temperature or cut with shears. Similarly, (17) will break down if we intervene to produce even a small extension in the wrong way—for, example, if the intervention physically deforms the spring. How we produce a given value of *X* in (17) and not just what that value is matter for whether (17) is invariant under that change. The set of possible 'interfering conditions' for (17) is very large and heterogeneous and will resist any simple, informative characterization. Because we don't know how to characterize all the items in this set in a way that is non-circular and illuminating, we find ourselves saying things like the following: (17) holds, if 'other things are equal' or in the absence of 'disturbing factors' where no very precise independent specification of the quoted phrases is available.

By way of contrast, although paradigmatic laws like Maxwell's equations do break down under certain extreme values of the variables that figure in those equations, whether the equations hold or not depends just on the values of those variables and not on how those values are brought about. That is, in contrast to (17), within the classical regime for which Maxwell's equations hold, it does not matter how we change the distances between point charges, or the intensities of electromagnetic fields and so on—Maxwell's equations will continue to hold under such changes. Moreover, changes in background conditions play a different rule in connection with the invariance characteristics of paradigmatic laws than in connection with generalizations like (17). When the circumstances under which paradigmatic laws fail to be invariant are known, they typically can be given a relatively simple, unified characterization. Such circumstances seem to fall into one of two categories: laws break down either for extreme values of variables that explicitly figure in the them (e.g. high temperatures and pressures in the case of the ideal gas law) or when some very small set of variables that have been omitted from the law diverge from a

limiting value—the pattern being that the law holds when the variables take this limiting value but not otherwise. For example, according to a well-known textbook on General Relativity, the Newtonian inverse square law 'is an excellent approximation in the limiting case of low velocity in a weak gravitational field' (Ohanian [1976], p. 2). That is, the law breaks down both when gravitational fields are strong (an extreme value of an included variable) and also when an omitted variable (velocity) is not small in comparison with the speed of light.

On this way of looking at matters, the differences between (17), on the one hand, and paradigmatic laws like Maxwell's equations, on the other, although real, look very much like differences in degree (of scope and of range of interventions and changes in background conditions over which these generalizations are invariant) rather than of kind. Paradigmatic laws are simply generalizations with wide scope that are invariant under a large and important set of changes that can be given a theoretically perspicuous characterization. We are willing to regard other invariant generalizations as laws to the extent that we judge that they resemble these paradigms in these respects. It is thus not surprising that the boundary between those invariant generalizations we regard as laws and those that we do not regard as laws is fuzzy and contentious—an additional reason for resisting models of explanation that require a sharp law/non-law boundary.

These considerations raise in turn an obvious question: given that the difference between Maxwell's equations and (17) is one of degree, why not reflect this continuity by extending the notion of a law to cover all generalizations that are invariant under some interventions and changes in background conditions, so that generalizations like (17) count as laws as well, albeit local or qualified or ceteris paribus laws? In fact, many writers have proposed that we do just this (Hausman [1992]; Fodor [1991]; Kincaid [1989]).

In thinking about this proposal, it is important to separate issues that are largely terminological in the sense that they reflect decisions about how to use the word 'law' from more substantive issues. To the extent that the proposal under discussion accepts what I have say about the importance of invariance and its role in explanation and simply extends the word 'law' to cover all invariant generalizations, it differs only verbally from my own position. In fact, however, few if any philosophers who have wanted to extend the notion of law have had in mind only such a terminological proposal. Instead, philosophers who have thought of generalizations like (17) (or the various generalizations of the special sciences) as laws have usually been motivated by a very different account from the one I have been defending of the features of such generalizations which make them explanatory. They have tried to show that such generalizations are explanatory in virtue of satisfying (or to the extent they satisfy) various of the traditional criteria for lawfulness rather than in

virtue of being invariant or figuring in the answers to a range of what-if-things-had-been-different questions. For example, the treatments of so-called *ceteris paribus* laws, which are discussed in Section **13**, are all motivated by the idea that if a generalization is to be a law and hence explanatory it must be or be backed by an exceptionless generalization. The philosophers who advocate such treatments are not merely proposing that we extend the word 'law' to cover the explanatory generalizations of the special sciences but are instead adopting a distinctive substantive position about the features (*viz*. exceptionlessness) which such generalizations must possess if they are to figure in explanations.

It is in part for this reason—that in practice, the project of extending the notion of law to cover the generalizations of the special sciences is closely bound up with the (in my view thoroughly misguided) project of showing that, despite appearances to the contrary, these generalizations are explanatory because they satisfy (at least many of) the traditional criteria for lawfulness—that I think that clarity is best served by adopting a more restricted notion of law. Moreover, there are additional reasons for such a restricted notion. First, while there are, as I have argued, important continuities between generalizations like Maxwell's and (17), there are also very real differences. While these may be matters of degree, they are not for that reason unimportant. The features possessed by generalizations, like Maxwell's equations—greater scope and invariance under larger, more clearly defined, and important classes of interventions and changes—represent just the sort of generality and unconditionality standardly associated with laws of nature. Their relative absence from generalizations like (17) and from many of the explanatory generalizations of the special sciences makes it misleading to assimilate these to paradigmatic laws. Second, and perhaps more importantly, if the argument of this paper about invariance and explanation is correct, there is no real motivation for such an assimilation. The claim that the explanatory generalizations of the special sciences are laws would have an obvious motivation if there was some independent reason for supposing that all explanation requires laws, understood along the traditional lines. It would also have an obvious motivation if there were some independent reason to suppose that all generalizations must fall into one of two mutually exclusive categories—the lawful or the purely accidental. However, I have argued that we should reject these assumptions. Both are gratuitous—we don't need to accept them once we have the notion of invariance and the account of explanation sketched above. Once we accept this alternative account of explanation, we don't need to argue that the generalizations of the special sciences are laws (thereby incurring the burden of claiming that they do not differ in important respects from Maxwell's equations or that, appearances to the contrary, they satisfy such traditional criteria as exceptionlessness) to vindicate their explanatory status. Finally, as

will became clear in Section **12**, the more restricted usage that I favor also has the advantage that it captures what seems to be at stake when philosophers and scientists deny, as they frequently do, that the explanatory generalizations of the special sciences are laws. Writers who take this position typically are not merely making a proposal about terminology; instead they think that there are important differences between generalizations like Maxwell's equations and many of the explanatory generalizations of the special sciences which a descriptively adequate account of explanatory practice in different areas of science should to aim to capture. The framework I have proposed allows us to do this.

12 Invariance and structural equations

I now turn to some examples drawn from the structural equations literature designed to illustrate in a more detailed way how the ideas about invariance and explanation developed in previous sections apply to generalizations in the special sciences.[10] *S*uppose that we are interested in determining the extent to which, for some population P of plants, the amount of water (X_1) and fertilizer (X_2) received by an individual plant in P influences its height *Y*. To this purpose we write down the linear regression equation

$$Y = a_1X_1 + a_2X_2 + U \tag{35}$$

Here a_1 and a_2 are fixed coefficients and *U* is a so-called error term, which we may take to represent other causal influences on *Y* besides X_1 and X_2 that have been omitted from (35).

What are the conditions that must be satisfied for (35) correctly to describe a causal or explanatory relationship? On my view, these are just the conditions set out in the general account of explanation in this essay: (35) must be invariant under some interventions (on the variables figuring in (35)) and must be such that it can be used to answer a range of what-if-things-had-been-different questions. In particular, it should be true for some interventions which change the right hand side variables X_i by the amount ΔX_i, that these also change *Y* in just the way represented by (35)—i.e. by $a_i\Delta X_i$. To express what is really the same idea in the language of invariance, if (35) correctly describes a causal or explanatory relationship, then the relationship represented by (35)—that is, its functional form and the values of the individual coefficients a_i—should be stable or invariant under some interventions on the right hand side variables. For example, an intervention on X_1 should not change the value of the coefficient a_1 or any of the other coefficients in (35), should not disturb the linear form of (35) and so on.

[10] For a more detailed application of these ideas to structural equations, see Hausman and Woodward ([1999]) and Woodward ([forthcoming]).

It should be clear that when (35) possesses this feature it will exhibit just the sort of pattern of active counterfactual dependence which we took to be crucial to successful explanation in Section **4** above. That is, when (35) is invariant under interventions in the way described we can think of it as answering a range of what-if-things-had-been-different questions about plant height—as telling us how the height of a plant (or perhaps the mean height in the population P) would change in various ways as the amount of water and fertilizer it receives is varied. In this way, we come to see how the amount of water and fertilizer a plant receives makes a difference for or is relevant to its height. Thus we can see in (35) the same features that we have found in the other examples of successful explanation discussed in previous sections. Obviously, (35) can be invariant in this way under some range of interventions even if (as is plainly the case) (35) breaks down under other interventions and background conditions, has limited scope, and lacks many of the other features traditionally assigned to laws of nature. As long as we are seeking to explain the behavior of a population of plants that falls within the domain of invariance of (35) we may legitimately appeal to it to explain, even though (35) fails to hold for other populations of plants in other background circumstances.

The generalization (35) concerns the behavior of plants but it is typical, in the respects just described, of many of the generalizations discovered by regression and other causal modeling techniques in social science contexts. Consider, as a second illustration, the investigation carried out by Eric Veblen in his ([1975]) and discussed at some length in Christopher Achen's illuminating monograph ([1982]). Veblen is interested in the effect of editorial endorsements by a particular newspaper, the *Manchester Union Leader*, on New Hampshire elections during the period 1960–72. He regresses a variable (vote difference) measuring the difference between the vote for the *Union Leader* candidate in Manchester (where the *Union Leader's* circulation is large) and the vote for this candidate outside of the Manchester area (where the newspaper's circulation is low) against a variable (slant) designed to measure the number of favorable news stories a candidate receives. Veblen finds that the *prima facie* effect of favorable coverage is quite large: a change from below to above average slant is associated with a 22 percent increase in the vote for the *Union Leader's* candidate.

What does it mean to claim, as Veblen and Achen do, that this relationship—call it (36)—is genuinely causal and that one can explain facts about the vote that various candidates receive by appealing to facts about the *Union Leader's* editorial policies? My suggestion is that what this means is that, given certain background conditions characteristic of New Hampshire during the period 1960–72, there is some range of interventions involving changes in editorial policies over which (36) will be stable or invariant—that it is not true that all such changes in editorial endorsements will disrupt (36)—and that,

because of this, the *Union Leader* can use its endorsements to change or manipulate voting patterns during this period. When the relationship between vote and slant is stable under changes in the amount of favorable editorial coverage a candidate receives in this way, one can appeal to such changes and this relationship to answer a range of what-if-things-had-been-different questions.

It is clear, however, that, like (35) and the Hooke's law relationship (17), the relationship (36) is, at best, stable or invariant only under a very limited range of interventions and changes in background circumstances. Veblen does not claim, and it is almost certainly not true, that either the precise quantitative relationship he finds or even the overall qualitative features of this relationship (that endorsements positively influence vote) would continue to hold across all changes in background conditions in the New Hampshire context or that these would hold for other newspapers in different circumstances. For example, if the extremely conservative *Union Leader* were to undergo a change in ownership and begin endorsing liberal candidates, it is doubtful that its endorsements would continue to have the same effect on the vote. Indeed, there are all sorts of possible changes in the political attitudes of New Hampshire voters, and the behavior of political parties and candidates within New Hampshire that would disrupt (36). Moreover, it is easy to imagine that in another population, the effect of favorable news coverage by the largest circulation local newspaper might be quite different. For example, the dominant newspaper in Southern California, the relatively liberal *Los Angeles Times*, is strongly disliked by many voters in neighboring conservative Orange County. One would guess that editorial endorsements and favorable news coverage by this newspaper would have a quantitatively smaller positive effect—if indeed the effect is positive at all—on the voting behavior of this electorate. Again the notion of invariance and the what-if-things-had-been-different-account of explanation give us a way of understanding how the relationship Veblen has discovered can be explanatory, despite the fact that it holds only a very restricted set of circumstances. As explained above, a relationship can be invariant even thought it has exceptions (outside of its domain of invariance), is very narrow in scope, holds only over a limited spatio-temporal interval and so on. We see all of these features in a relationship like (36).

The idea that, even when causal, the relationships discovered by regression and other causal modeling techniques are likely to be invariant across only very limited set of interventions and changes in background conditions or across restricted spatio-temporal intervals is widely, although far from universally acknowledged, by social scientists themselves. For example, Achen ([1982], p. 12) uses Veblen's research to motivate the more general claim that usually in social science:

> the researchers' intent is to describe [. . .] for example [. . .] the effect of the Catholic vote on the Nazi rise to power, or the impact of a preschool

> cultural enrichment program like Head Start on poor children's success in school. *Whatever the truth in such cases, one would not characterize it as a law. Neither Catholics nor impoverished youngsters would behave the same way in other times and places* (my emphasis).

Elsewhere, he adds that in social science, 'functionally specific laws are sure to fail serious empirical tests. They always have' ([*ibid.*], p. 15).

When Achen claims that relationships like (36) are explanatory but denies that they are laws, what he seems to have in mind is precisely the sort of contrast described in section **10** between laws and other sorts of invariant relationships. Generalizations like (36) are so restricted in scope and range of invariance that it strikes Achen as unilluminating and misleading to assimilate them to paradigmatic laws like Maxwell's equations. On the other hand, one can nevertheless appeal to these generalizations to explain. The notion of invariance and the what-if-things-had-been-different account of explanation show how this is possible.

Before proceeding, let me acknowledge a natural response to generalizations like (35) and (36). This is that although they are not entirely unexplanatory, they are shallow and superficial. One doesn't understand at a very deep level why the plants in P grow as they do if one only knows (35). A deeper explanation of why the plants grow to the height they do would require a much more detailed understanding of plant development and metabolism and of the biochemical and physiological mechanisms by which water, fertilizer, and other factors influence growth. Similarly in the case of (36), a deeper explanation of the behavior of the New Hampshire electorate would focus (among other things) on the psychological mechanisms by which editorial endorsements influence voting behavior.

I fully endorse this assessment of (35) and (36) but see it as supporting rather than undermining the ideas about invariance and explanation developed above. The generalizations—call them (37)—describing the biochemical, molecular and cellular processes underlying plant growth do not provide information that is different in kind from the information provided by (35) but rather provide, as it were, more of the same. The generalizations (37) will be invariant under a wider range of interventions and changes than (35) and they can be used to answer a wider range of what-if-things-had-been-different questions. For example, we might plausibly hope that these generalizations will continue to hold if we apply amounts of water and fertilizer for which the relationship (35) breaks down and hence can be used to explain why such departures from linearity occur. Thus the relationship between (35) and (36) will be like the relationship between the shallow gas pedal-speed relation (13) and the deeper engineering theory of automotive mechanics in Haavelmo's example (Section **5**) and should be understood along similar lines.

13 Invariance and *ceteris paribus* laws

It will help to bring out what is distinctive about the ideas about invariance and explanation that I have been defending if we contrast them with a standard alternative account of the conditions that explanatory generalizations in the special sciences must meet. I will call this the *completer* account—versions can be found in many writers, from Hempel ([1965]), to more recently Fodor ([1991]), Hausman ([1992]) and Pietroski and Rey ([1995]). Each of these writers adds refinements and complications, but in what follows I will focus on the core idea. I shall argue that this is sufficiently mistaken that the embellishments will not help. The completer account adopts the assumptions with which we began this paper—that all explanation requires 'subsumption' under laws and that a necessary condition for a generalization to count as a 'strict' or unproblematic law is that it be exceptionless. (Recall the passage from Pietroski and Rey (1995), quoted in Section **8**.) We then face the familiar problem of reconciling these assumptions with the apparent paucity of exceptionless generalizations in the special sciences. In schematic form, the solution proposed by the completer account is this: suppose one begins with a generalization of form (37) 'All *Fs* are *Gs* ' which has exceptions. (37) will be a legitimate kind of law—a so-called *ceteris paribus* law—and will have explanatory import if and only if there is some further condition *C* such that (38) 'All *Fs* in C are *Gs*' is a strict or exceptionless law (that is, *F* and *C* is nomologically sufficient for *G*) and neither *F* by itself nor *C* by itself is nomologically sufficient for *G*. Adopting (and somewhat modifying) some terminology due to Fodor ([1991]), let us call such a condition *C* a 'completer' for (37). The simplest version of the completer account then says that a ceteris paribus generalization is genuine law and hence explanatory if and only if it has a completer. It is crucial to the structure of this account that we *not* impose the requirement that some one who appeals to (37) to explain must be able to actually describe its completer *C* in a non-trivial way or to state or produce the exceptionless generalization (38). As we have repeatedly noted, it is rarely possible to do this in the special sciences. Instead, it is enough that the completer exists or, alternatively, perhaps that we know or have some reason to think that it exists, even if we are unable to provide a non-trivial description of it.

The apparent attraction of the completer strategy is that it allows one to retain the idea that there is a sense in which explanation requires laws and that laws must be exceptionless while at the same time according an explanatory role to generalizations that have exceptions—this is accomplished by requiring that (37) be 'backed' or associated with the exceptionless law (38). The idea is that somehow (38), in virtue of its exceptionlessness, endows (37) with explanatory import and gives it a status as a legitimate (*ceteris paribus*) law that it would not

possess if it did not have a completer.[11] We can also think of this strategy as illustrating a general point made in Section **11**—that the claim that a generalization like (37) is a 'law' (albeit a *ceteris paribus* law) becomes a substantive claim and not merely a recommendation about terminology when it is embedded in a more general set of ideas about the features a generalization must possess (in this case, completability into an exceptionless generalization) if it is to be explanatory. The claim that (37) is explanatory when and only when it has a completer is a substantive claim, not a bit of verbal stipulation.

Despite its apparent naturalness, I believe that the completer strategy is fundamentally flawed and that an appreciation of these flaws will bring out the superiority of the invariance based account that I favor. To begin with, the underlying motivation for the account is problematic. This motivation depends on the idea that there is an invidious contrast between *ceteris paribus* laws like (37), which have exceptions, and genuine or strict laws which are exceptionless, and that, because of this, to vindicate the former we must show that they are backed in an appropriate way by the latter. However, as we have already observed and as defenders of the completer account readily acknowledge, there are few examples of exceptionless laws to be found anywhere in science, even in fundamental physics. (Indeed, both Fodor ([1991]) and Pietroski and Rey ([1995]) explicitly say that there may be *no* known examples of exceptionless laws.) Surely, the natural conclusion to draw from this observation is that the whole idea that genuine laws must be exceptionless and that explanation requires exceptionless laws needs rethinking. It is just this conclusion that I have advocated in this essay.

[11] There is an epistemic puzzle here. Consider first the simplest version of the completer account according to which the mere existence of a completer endows a generalization with explanatory import even if those who use the generalization are unaware of its existence. It is far from obvious how this is possible. In my view it is a plausible principle that explanatory information must be information that is epistemically accessible. Explanations work by conveying information that provides understanding and this seems to mean that such information must be epistemically accessible to those who provide the explanation or are enlightened by it. If some of the information allegedly conveyed by an explanation is information that cannot be grasped or recognized by those who use the explanation, we may reasonably doubt that it is through recognition of that information that the explanation provides understanding. If no one knows that (33) has a completer (or even what a completer is) and yet (33) is used to explain, it is hard to believe that its explanatory import is crucially bound up with whether it has a completer. I think that a parallel conclusion holds if it is instead claimed that for (33) to be explanatory those who appeal to it or those in the audience to which it is directed must know that it has a completer, even if they do not know the identity of the completer. Why should merely knowing that (33) has a completer but not the identity of the completer endow it with explanatory import? By contrast, the account I have developed makes explanatory import depend only on epistemically accessible information. Typical users of explanations of the sort described in previous sections are able to recognize that they are within the domain of invariance of the generalizations they employ and are able to see how these generalizations can be used to answer a range of what-if-things-had-been-different questions. Information that is typically not epistemically accessible to users, such as information about the exact boundaries of domains of invariance or the full range of circumstances in which a generalization will break down is not information that is needed in order to successfully explain.

In fact, as I shall now argue, the fundamental intuition underlying the completer strategy is wrong—the distinction between those generalizations that have completers and those that do not does not coincide with the distinction between those generalizations that are lawful (or invariant or explanatory) and those that are not. Under the assumption of macrodeterminism, which virtually all defenders of the completer strategy endorse, there are many generalizations that have completers that no one would regard as explanatory or as ceteris paribus laws. Consider the generalization

(39) All human beings with normal neurophysiological equipment speak English with a southern U. S. accent.

This generalization is of course false—it has exceptions—but under the assumption of determinism there will be a very complicated set of conditions *K* that are nomologically sufficient in conjunction with being a human being with a normal neurophysiology for speaking English with a southern accent and which satisfy the other conditions for being a completer, such as those described in Pietroski and Rey ([1995]). Indeed, we even have a general sense of what those conditions are—they include some very complex set of environmental conditions including appropriate early exposure to English spoken with a southern accent. These together with being a human being with the appropriate neurological structures are nomologically sufficient to insure that one will learn to speak English with a southern accent. So (39) has a completer—(40) 'All human beings with normal neurophysiology in *K* speak English with a Southern accent' is not just exceptionless but arguably satisfies many of the other standard conditions for lawfulness such as support for counterfactuals. Nonetheless, (39) is surely not the kind of generalization that anyone would regard as a ceteris paribus law—and not just because it has many exceptions. Even if we imagine that (39) is exceptionless—that, as the result of political and economic changes, all living humans were to come to speak English with a southern accent and even if past history had been different in such a way that throughout all history, all human beings spoke English with a southern accent—(39) would not be a plausible candidate for a law of any sort, *ceteris paribus* or otherwise. Nor can we appeal to (39) to provide an explanation of why some particular person speaks English with a southern accent.[12] (We can, of course, appeal to (40) to explain this but (40) is not (39).)

By contrast, the invariance-based account does explain in a natural way why (39) is a poor candidates for an explanatory generalization—it is either non-invariant or invariant only under a very narrow range of interventions. Even if as a result of political changes it becomes true at some future date that everyone

[12] A natural question is whether there are plausible additional constraints which when added to the version of the completer account described above will avoid these difficulties. While I cannot canvass all of the possibilities, some brief remarks on some of the additional constraints already

in the world speaks English with a southern accent and even if, because the past history of the human race was very different, it was true that all human beings in the past had spoken southern English, (39) would still be highly fragile. Its truth would depend upon a great many very specific contingencies and if these were to change, (39) would be disrupted. Only in a very special and rare kind of environments (rare in comparison with the full range of environments in which human beings learn to speak some language or other) do human beings learn to speak English with a southern accent.

We can further bring out the difference between the completer account and the invariance based account by considering what the former has to say about the contrast between, on the one hand, the 'shallow' generalization (13) linking the position of the gas-pedal and the speed of a car and, on the other, the deeper engineering style theory (14) described in Haavelmo's example (Section **5**). (13) has exceptions but, according to the completer strategy, will qualify as a ceteris paribus law because there exists some complicated condition *K* (specifying the details of the functioning of the car engine and the environmental

suggested in the literature will help to support my claim that the difficulties are not easily avoided. Pietroski and Rey ([1995], p. 90) add the additional constraint that the completer must be 'independent' where this is understood in such as way as to 'exclude factors whose only explanatory role is to save a proposed c. p. law' and to require that the completer must explain other things besides the failure of the law. Their discussion conflates two distinct issues: whether the completer is independent in the sense just specified and the epistemic issue of whether one knows how to independently describe the completer in a way that is not tantamount to saying that the law holds except when it doesn't. The examples described above do involve completers that are independent in the non-epistemic sense described in the quoted passage and hence are counterexamples to their proposal. Hausman ([1992], p. 137) claims that 'when one takes an inexact generalization to be an explanatory law, one supposes that the *ceteris paribus* clause picks out some predicate that when added to the antecedent of the unqualified generalization makes it an exact law'. This formulation appears to be straightforwardly subject to the counterexamples given above. Hausman ([*ibid.*], pp. 140ff) also describes an additional set of conditions which must be satisfied for it to be reasonable to believe that an inexact *ceteris paribus* generalization is completable into an exact law. The conditions appear to be unnecessary if, as I have argued, the thesis of macrodeterminism by itself gives one very general reasons to believe that there must be a completer. Moreover, quite apart from what it is reasonable to believe, the fact remains that (39) is a true *ceteris paribus* law, according to Hausman's characterization, given the empirical assumptions described above and this is surely an unwelcome result. Would it help if we instead took Hausman's additional conditions as necessary conditions for it to be true that a generalization qualifies as a *ceteris paribus* law rather than as conditions on what we have reason to believe? However we understand the role of the conditions, several of them seem wrong-headed. For example, one of the conditions is that a ceteris paribus law must be 'reliable'. This requires that it be true, for all *Fs* are *Gs* to be a *ceteris paribus* law, that '(perhaps after making allowances for specific interferences), almost all *Fs* are *Gs*' (p. 140). But, as explained above even if (39) were true without exception, it wouldn't be a *ceteris paribus* law. Another of Hausman's conditions is that the 'modifications or qualifications of the theory that make [a candidate *ceteris paribus* generalization] more reliable not be *ad hoc*' (p.141). But whether the completer for a candidate c. p. law is simple or complex, *ad hoc* or non-*ad hoc*, appears to have little to do with whether it is invariant at all, and if so, over what range of interventions, and hence little to do with whether it can be used to explain. Many generalizations in the special sciences probably have very complex and *ad hoc* completers but this fact by itself does not make them unexplanatory. Conversely, it seems perfectly possible for a generalization to have a non-*ad hoc* completer and yet be non-invariant and non-explanatory—again (39) may be a case in point.

conditions in which it is operated) such that the generalization 'In *K*, (13) holds' is an exceptionless law. For similar reasons, (14) will also qualify as a non-strict, *ceteris paribus* law—there will be circumstances in which it breaks down but it will also have a completer. However, the invariance-based approach allows us to say that (14) furnishes a deeper explanation of the behavior of the car because (14) is invariant under a wider range of interventions than (13) and can be used to answer a wider range of what-if-things-had-been-different questions. By contrast, the complete strategy provides no basis for such a discrimination—all that it says about (13) and (14) is that they are both *ceteris paribus* laws. Again, I take this to illustrate how the invariance-based account focuses on a very different set of considerations in assessing the explanatory credentials of a generalization than the completer strategy. Simply asking whether a generalization has a completer gives us no insight into the range of changes over which it is invariant and the range of what-if-things-had-been-different questions it can be used to answer. Yet it is just these questions that are crucial to explanatory assessment.

Both of these examples illustrate the more general point that when we ask whether a generalization has a completer and when we ask whether it is invariant we are asking very different questions. The invariance of a *ceteris paribus* generalization like (37) 'All *Fs* are *Gs*' depends not just on whether it has a completer *C* but on the details of the way in which the holding of (37) depends on both on *C* and on the various alternatives to *C*—on whether, for example, *C* represents a special case with (37) holding only when *C* does or whether, on the other hand, *C* is a more generic case and (37) would continue to hold under some range of alternatives to *C*. Again, the relevance of these sorts of considerations to explanatory assessment are lost if we focus only on the question of whether (37) has a completer.

13 Conclusion

In this paper I have argued for a number of general claims. First, explanations involve the exhibition of patterns of active or non-backtracking counterfactual dependence rather than nomic subsumption. Second, explanatory generalizations describe how changes in a set of explanans variables produce changes in a set of explanandum variables and must be invariant under some range of interventions on the explanans variables. Third, the requirement that explanatory generalizations must be invariant is very different from the traditional demand that explanatory generalizations must be laws. Unlike lawfulness, invariance admits of degrees. Moreover, the traditional criteria for lawfulness are neither necessary nor sufficient for a generalization to be invariant or explanatory. In particular, a generalization can be explanatory even if it is not exceptionless and is very restricted in scope. A focus on invariance leads to a

more illuminating account of the nature of explanatory generalizations in the special sciences than alternatives such as the completer account.

Acknowledgements

Thanks to Nancy Cartwright, Lindley Darden, Dan Hausman, Chris Hitchcock, Paul Humphreys, Marc Lange, and Judea Pearl for helpful conversations and comments and to three anonymous referees for constructive criticisms. Research for this paper was supported by a grant from the National Science Foundation (SBR-9320097).

Division of Humanities and Social Sciences
California Institute of Technology
Pasadena, CA 91125
USA
jfw@hss.caltech.edu

References

Achen, C. [1982]: *Interpreting and Using Regression*, Beverly Hills: Sage Publications.

Beatty, J. [1979]: 'Traditional and Semantic Accounts of Evolutionary Theory', unpublished Ph.D. dissertation, Ann Arbor, MI: University Microfilms International.

Cartwright, N. [1983]: *How the Laws of Physics Lie*, Oxford: Clarendon Press.

Dunn, L. C. [1957]: 'Evidence of Evolutionary Forces Leading to the Spread of Lethal Genes in Wild populations of House Mice', *Genetics*, 43, pp. 157–63.

Fiorina, M. [1995]: 'Rational Choice, Empirical Contributions and the Scientific Enterprise', in J. Friedman (*ed.*), *The Rational Choice Controversy: Economic Models of Politics Reconsidered*, New Haven: Yale University Press.

Fodor, J. [1991]: 'You Can Fool Some of the People All of the Time, Everything Else Being Equal: Hedged Laws and Psychological Explanation', *Mind*, **100**, pp. 19–34.

van Fraassen, B. C. [1989]: *Laws and Symmetry*, Oxford: Clarendon Press.

Friedman, J. (*ed.*) [1995]: *The Rational Choice Controversy: Economic Models of Politics Reconsidered*, New Haven: Yale University Press.

Friedman, M. [1974]: 'Explanation and Scientific Understanding', *Journal of Philosophy*, **71**, pp. 5–19.

Giere, R. [1988]: *Explaining Science: A Cognitive Approach*, Chicago: University of Chicago Press.

Green, D. and Shapiro, I. [1995]: 'Reflections On Our Critics, in J. Friedman (*ed.*), *The Rational Choice Controversy: Economic Models of Politics Reconsidered*, New Haven: Yale University Press.

Haavelmo, T. [1944]: 'The Probability Approach in Econometrics', *Econometrica*, **12** (Supplement), pp. 1–118.

Hausman, D. [1992]: *The Inexact and Separate Science of Economics*, Cambridge: Cambridge University Press.

Hausman, D. and Woodward, J. [1999]: 'Independence, Invariance and the Causal Markov Condition', *British Journal for the Philosophy of Science* **50**, pp. 521–83.

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

Hitchcock, C. and Woodward, J. [unpublished manuscript]: 'Explanatory Generalizations Deep and Shallow'.

Johnston, J. [1992]: 'Econometrics: Retrospect and Prospect', in J. Hey (*ed.*), *The Future of Economics*, Oxford: Blackwell.

Kincaid, H. [1989]: 'Confirmation, Complexity and *S*ocial Laws', in A. Fine and J. Leplin (*eds*), *PSA 1988*., East Lansing, MI: Philosophy of Science Association.

Kitcher, P. [1989]: 'Explanatory Unification and the Causal Structure of the World', in P. Kitcher and W. Salmon, *Scientific Explanation*, Minneapolis: University of Minnesota Press, pp. 410–505.

Lewis, D. [1973a]: *Counterfactuals*, Cambridge: Harvard University Press.

Lewis, D. [1973b]: 'Causation', *Journal of Philosophy*, **70**, pp. 556–67. Reprinted with Postscripts in Lewis [1986], pp. 159–213.

Lewis, D. [1986]: *Philosophical Papers, Vol. II*, Oxford: Oxford University Press.

Lewontin, R. and Dunn, L. [1960]: 'The Evolutionary Dynamics of a Polymorphism in the House Mouse', *Genetics* , **45**, pp. 705–22.

Lucas, R. [1983]: 'Econometric Policy Evaluation: A Critique', in K. Brunner and A. Meltzer (*eds*), *Carnegie–Rochester Conference on Public Policy, Supplementary Series to the Journal of Monetary Economics*, **1**, Amsterdam: North Holland, pp. 257–84.

Luce, R. D. and Raiffa, H. [1957]: *Games and Decisions*, New York: John Wiley & Sons.

Lyon, A. [1977]: 'The Immutable Laws of Nature', in *Proceedings of the Aristotelian Society 1976–77*, London: Compton Press, pp. 107–26.

Ohanian, H. [1976]: *Gravitation and Spacetime*, New York: W. W. Norton & Company.

Pietroski, P. and Rey, G. [1995]: 'When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity', *British Journal for the Philosophy of Science*, **46**, pp. 81–110.

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

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

Satz, D. and Ferejohn, J. [1994]: 'Rational Choice and Social Theory', *Journal of Philosophy*, **91**, pp. 71–87.

Veblen, E. [1975]: *The Manchester Union-Leader in New Hampshire Elections*, Hannover, NH: University Press of New England.

Woodward, J. [1979]: 'Scientific Explanation', *British Journal for the Philosophy of Science*, **30**, pp. 41–67.

Woodward, J. [1984]: 'A Theory of Singular Causal Explanation', *Erkenntnis*, **21,** pp. 231–62. Reprinted in D. Reuben (*ed.*), *Explanation*, Oxford: Oxford University Press [1993], pp. 246–74.

Woodward, J. [1997a]: ‘Explanation, Invariance and Intervention’, in *PSA 1996,* **2**, pp. 26–41.

Woodward, J. [forthcoming]: ‘Causal Interpretation in Systems of Equations’, *Synthese.*