

## Explanatory Generalizations, Part I: A Counterfactual Account\*

### 1. Introduction

The nomothetic conception of explanation, according to which all successful explanations must appeal to laws, has dominated the discussion of scientific explanation in the second half of the twentieth century. The best known formulation of the nomothetic conception of explanation is, of course, Hempel's Deductive-Nomological theory of explanation. While few philosophers today accept the D-N theory of explanation in its original formulation, there is a widespread consensus that laws play a central role in explanation, even among prominent critics of the D-N model such as Wesley Salmon (see, e.g., Salmon 1984, p. 262).

This emphasis on the role of laws naturally raises the question: 'what is a law of nature?' The standard answer is that laws are (or at least entail) exceptionless generalizations. Not all exceptionless generalizations are laws, however. It may be that all of the members of the Greenspoint School Board are bald, but this is not a generalization that could be used to explain why some individual member of the school board is bald. So what more is needed? Various other conditions have been proposed: laws must contain only qualitative predicates, support counterfactuals, be confirmed by their instances, and so on. However, there is general agreement that none of these proposals (either singly or in combination) is completely successful (see, for example, the discussion in Salmon's (1989) survey). Defenders of the nomothetic conception are thus in the uncomfortable position of insisting that laws are essential to successful explanation while lacking a clear account of what laws are.

The assumed role of laws in explanation also gives rise to a related problem. Fields of scientific inquiry that deal with complex systems — the life sciences and social sciences, as well as branches of the physical sciences such as meteorology and geology — seem to provide generalizations that are not truly exceptionless and which lack many of the other features standardly assigned to laws. John Beatty, in an interesting series of articles, has argued on these grounds that there are no laws of biology. (See, e.g., Beatty 1995; see also Smart 1963.) Others have argued for similar conclusions in the social sciences (see, e.g., Scriven 1956, Fay 1983, Rosenberg 1992, Earman and Roberts 1999). This presents us with an apparent dilemma:

[If] one insists that the special sciences don't state laws, one must either (a) explain...how the special sciences can provide good explanations without having laws to avert to, or (b) deny the immensely plausible claim that...the special sciences sometimes provide good explanations. (Pietroski and Rey 1995, 85.)

Pietroski and Rey apparently regard the first horn as so unattractive that they do not explore it further; their unwillingness to do so is a testament to the lasting influence of the nomothetic conception of explanation.

In this paper we will defend a new theory of explanation and explanatory generalizations that attempts to meet the challenge described under (a). The central idea is that successful explanation has to do with the exhibition of patterns of counterfactual dependence describing how the system whose behavior we wish to explain would change under various conditions. As we will see, whether a generalization can figure in such a pattern of dependence and hence 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 would continue to hold under an appropriate class of changes involving interventions on the variables figuring in that generalization. Invariant generalizations (and only invariant generalizations) will support the kinds of counterfactuals required for successful explanation. Many of the generalizations of the special sciences are invariant and hence explanatory even though they are not naturally regarded as laws. While the absence of a generally accepted account of lawfulness is a serious problem for the nomothetic conception, our theory

avoids this problem because it does not rely on the notion of law but rather on the notion of invariance, which, we will argue, is a relatively clear notion.

Our discussion is organized as follows. Section two motivates the claim that explanation has to do with the exhibition of patterns of counterfactual dependence. Section 3 provides a more detailed explication of the concepts of invariance and intervention. In section 4, we compare our account of explanation with Hempel's D-N account. Here we will show that our account differs from the D-N model in the kind of generality it takes to be desirable in explanations. While the D-N and other traditional models of explanation emphasize generality with respect to objects or systems other than the one that is the focus of explanation, our account instead emphasizes generality with respect to other possible properties of the very object or system that is the focus of explanation. We argue in a companion piece to this paper (Hitchcock and Woodward forthcoming — hereafter referred to as EG2) that a focus on this second sort of generality permits a much more satisfying account of what it means to say that one explanation is deeper than another.

Before turning to details, we should remark that there is an alternative way of conceiving of our project. Readers who wish to retain the idea that laws are essential to explanation may view our account of explanatory generalizations as a new account of laws, rather than as an argument that generalizations that are not laws may figure in explanations. The difference between this perspective and our own strikes us as largely verbal. The substantive point on which we disagree with philosophical tradition has to do with features that generalizations must possess if they are to play an explanatory role: we think it is invariance, rather than exceptionlessness or any of the other features traditionally associated with laws, that is crucial. What we decide to call explanatory generalizations is secondary. However, some convention about how to use the word 'law' is required, and we think that it is simplest and least confusing to restrict the word 'law' to its traditional philosophical meaning.

## 2. Of Laws and Explanations

According to Hempel's D-N theory of explanation, explanations have the following logical form:

$$\begin{array}{l} \underline{C}_1, \underline{C}_2, \dots, \underline{C}_m \\ \underline{L}_1, \underline{L}_2, \dots, \underline{L}_n \\ \therefore \underline{E} \end{array}$$

E is a proposition describing the phenomenon to be explained — the explanandum. E is derived from a set of other propositions, collectively called the explanans. The explanans contains propositions of two distinct types: C<sub>1</sub>, C<sub>2</sub>, ... C<sub>m</sub> describe particular circumstances or initial conditions, while L<sub>1</sub>, L<sub>2</sub>, ... L<sub>n</sub> describe laws of nature. The laws of nature must figure essentially in the derivation; the derivation is invalid without these premises. When such a derivation is given, it shows that the explanandum was to be expected in light of the explanatory information. It is for this reason, according to Hempel, that D-N explanations explain.

Many explanations conform to this structure. Consider an explanation of the magnitude of the electric field created by a long, straight wire with a positive charge uniformly distributed along its length. A standard textbook account proceeds by modeling the wire as divided into a large number of small segments, each of which acts as a point charge of magnitude dq. Each makes a contribution dE to the total field E in accord with a differential form of Coulomb's law:

$$(1) \quad \underline{dE} = (1 / 4\pi\epsilon_0)(\underline{dq} / \underline{s}^2)$$

where s is the distance from the charge to an arbitrary point in the field. Integrating over these individual contributions yields the result that the field is at right angles to the wire and that its intensity is given by

$$\underline{E} = (1 / 2\pi\epsilon_0)(\lambda / \underline{r})$$

where  $r$  is the perpendicular distance to the wire and  $\lambda$  the charge density along the wire.

This explanation does instantiate the D-N schema: it consists of a deductively valid argument in which a law of nature, in this case Coulomb's law, figures as an essential premise. However, we want to focus on an additional feature that plays no role in the D-N model. Put abstractly the feature is this: the generalization (1) not only shows that the explanandum was to be expected, given the initial conditions that actually obtained, but it can also be used to show how this explanandum would change if these initial and boundary conditions were to change in various ways. As we will put it, (1) can be used to answer a range of what-if-things-had-been-different questions. For example, (1) can be used to tell us how the electric field would differ if the charge density of the wire were increased, or if the wire twisted into a circle or a solenoid. In this way, (1) shows us that certain factors, such as the charge density and geometrical configuration of the conductor, make a systematic difference to the intensity and direction of the field. In short, Coulomb's law is explanatory because it tells us what the electric field depends on.

Based on this one example, of course, it is hard to adjudicate between these two rival analyses of what makes Coulomb's law explanatory. Both of us have defended elsewhere the idea that tracing dependence relations in the manner described above is essential to explanation (Woodward 1979, 1984, 1997a, 2000; Hitchcock 1993, 1995). While we will not rehearse this defense here, we will present one line of argument that will help to motivate the central idea.

The generalization (1) is commonly regarded as a law of nature. However, generalizations that are not plausibly regarded as laws also figure in explanations. Such generalizations can also be used to answer a range of what-if-things-had-been-different questions, just as Coulomb's law can. It is for this reason, we claim, that they also can be used to provide explanations. It is because there are generalizations and patterns of argument that answer such questions without citing laws that non-lawful explanation is possible.

Consider an illustration drawn from the structural equations literature. Suppose that we are interested in determining the extent to which the amount of water ( $X_1$ ) and fertilizer ( $X_2$ ) received by an individual plant influences its height  $Y$ . To this purpose we write down the linear regression equation

$$(2) \ Y = a_1 X_1 + a_2 X_2 + U$$

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  $X_2$ .

Even if the generalization (2) conveys information about a causal relationship between  $X_1$ ,  $X_2$  and  $Y$ , it falls far short of the standards normally demanded of laws. To begin with, (2) is bound to fail for sufficiently large or extreme values of  $X_1$  and  $X_2$ . One can't make a plant grow arbitrarily high by dumping huge amounts of water and fertilizer on it, nor make it arbitrarily small by giving it only minuscule amounts. Even if we confine our attention to values of  $X_1$  and  $X_2$  within a more ordinary range, (2) may fail to hold for certain ways of achieving those values; it will fail for instance, if  $X_1$  has a relatively high value as a result of dumping a large amount of water on the plant(s) at the end of an otherwise dry growing season. In addition, there are many background conditions, not represented in (2), which if changed would disrupt (2). (2) would fail if we were to spray the plant with weed killer or heat it to a very high temperature. Less dramatically, there are many possible conditions that will not destroy the plant, but which will alter the effect of water and fertilizer on plant height. There may be physical changes in the root system of the plant or the surrounding soil that would change the way in which given amounts of water affect plant height. Finally, even when we confine ourselves to the actual background conditions, and moderate quantities of water and fertilizer, (2) may not perfectly describe the relationship between the variables in question. While these features of (2) may make us reluctant to describe it as a law of nature, we commonly take such generalizations to provide useful information about

which variables are causally or explanatorily relevant to which others. (For a similar observation, see Earman and Roberts 1999, 12.)

What conditions must an equation like (2) satisfy if we are to regard it as making a true causal claim? In particular, how do we distinguish between the use of an equation like (2) to describe or summarize patterns of covariation among variables within a body of data, and its use to make causal claims or to explain? Consider the following passage from the statistician David Freedman:

Causal inference is different [from descriptive summary], because a change in the system is contemplated: for example, there will be an intervention. Descriptive statistics tell you about the correlations that happen to hold in the data: causal models claim to tell you what will happen to  $\underline{Y}$  if you change  $\underline{X}$ . (Freedman 1997, p.116)

We take Freedman's idea to be this: if (2) correctly describes a causal or explanatory relationship, then an intervention on the right hand side variables which changes each  $\underline{X}_i$  by the amount  $\Delta X_i$  should change  $\underline{Y}$  in just the way represented by (2) — i.e. by  $a_1 \Delta X_1 + a_2 \Delta X_2$ .<sup>1</sup> This is just to say that (2) should be invariant under interventions on the right hand side variables. When (2) has these properties it can be used to answer a range of what-if-things-had-been-different questions about how the height of the plant would change as the amount of water and fertilizer it receives is (hypothetically) varied. Thus we can see in (2) the same features that we have found in (1). Our claim is that we may regard (2) as explanatory in virtue of its possessing these features, even if (2) falls short of the standards we require in laws of nature.

We hope that this informal discussion conveys some sense of what it means to say that a relationship is invariant under interventions — we will offer a more precise characterization in section 3 below. We can use this notion to give a provisional formulation for our theory of explanation which will help to guide the reader through our subsequent discussion. An explanation involves two components, the explanans and the explanandum. The explanandum is a true (or approximately true) proposition to the effect that some variable (the 'explanandum variable') takes on some particular value. The explanans is a set of propositions, some which specify the actual (approximate) values of variables (explanans variables), and others which specify relationships between the explanans and explanandum variables. These relationships must satisfy two conditions: they must be true (or approximately so) of the actual values of the explanans and explanandum variables, and they must be invariant under interventions.

Note the similarity in structure between this formulation and the formulation of Hempel's D-N theory of explanation. The statement specifying the value of the explanandum variable is analogous to Hempel's explanandum proposition; the statements specifying the values of the explanans variables are analogous to Hempel's initial conditions; and the invariant generalizations figuring in our explanans are analogous to Hempel's laws. This similarity of structure will help to bring the essential differences between the two accounts into sharper focus. We will explore these differences in greater detail in section 4 below.

Two further clarifications are in order. First, our account can be extended analogously to include invariant generalizations relating the values of variables to the probability of some outcome, or the probabilistic distribution or expectation of some variable. Consider the following generalization: if a photon which is polarized with an angle of  $\theta_1$  from the vertical impinges on a polarizer set at angle  $\theta_2$  from the vertical, the probability that it will pass through is  $\cos^2(\theta_2 - \theta_1)$ . This generalization describes an invariant relationship between the probability of transmission and angular displacement and hence is potentially explanatory.<sup>2</sup>

Second, we will restrict our attention to causal explanation and will use the words 'causes' and 'explains' interchangeably in what follows. Perhaps some generalization of the account presented here can be developed for explanation in fields like mathematics and linguistics — we will return to this possibility briefly in EG2. We will also avoid cases in which the explanandum event is brought about in some idiosyncratic way, such as by preempting or overdetermining causes. We

think that the general treatment of causation in terms of counterfactual dependence that we favor can be extended to cover such cases, but we will not attempt to show this here.<sup>3</sup>

### 3. Invariance and Interventions

Obviously, our account puts a great deal of weight on the concepts of invariance and intervention, and it is to these concepts that the current section is devoted. A relationship is invariant if it continues to hold, or rather would continue to hold, in the presence of a certain range of changes. Our purpose in this section, then, is to specify the sorts of changes under which an explanatory relationship must remain invariant.

Consider a standard example of a relationship that does not carry explanatory import. Let  $\underline{X}$  represent the height of a column of mercury in a particular barometer located at Burbank Airport, and  $\underline{Y}$  the amount of rainfall recorded during a certain period at the same location. Then we might well expect there to be some relationship of the form

$$(3) \quad \underline{Y} = f(\underline{X}) + \underline{U},$$

that accurately predicts the amount of rain to be expected given any level of mercury in the barometer, where  $\underline{U}$  is as before an error term representing omitted causes. If our account is to be adequate, (3) had better not count as being invariant in the relevant sense.

There are in fact, at least two distinct sorts of changes under which (3) will continue to hold. We need to distinguish between stability under such changes and the sort of stability which establishes invariance. First, (3) will hold under changes in the price of tea in China, under changes in the lemming population of Norway, and under a great many other changes beside. These all involve changes in background conditions, that is, in the values of variables that do not explicitly figure in the generalization (3). (Of course, (3) will also break down under some changes in background conditions — for example, under extreme changes in temperature. However, since we require only that an explanatory relationship remain invariant under some range of changes, the existence of some changes in background conditions under which (3) will not continue to hold does not by itself show that (3) is non-invariant.) By contrast, the notion of invariance that we take to be central to explanation is invariance under some range of changes in the variables figuring in the relationship itself. This focus is one important respect in which our account differs from other accounts in the philosophical literature which appeal to ideas resembling our notion of invariance. For example Skyrms' (1980) notion of 'resilience' and Eells' (1991) requirement of 'context-unanimity' for probabilistic causation both incorporate the idea that causal and/or explanatory relationships satisfy a kind of stability condition, but this is understood as stability across background conditions. Unlike our view, these authors assign no special status to stability under interventions on the variables figuring in the relationship itself.

To count as invariant, however, it is not enough that a generalization remain valid under changes in the variables that figure in the relationship, for there is a natural sense in which (3) continues to hold under such changes: as we observe the mercury rise and fall as a result of normal atmospheric changes, we will observe that (3) continues to hold. To put the point another way, there are two different ways of completing the counterfactual: if the height of the column of mercury had been different, then... One might reason: if the height of the column of mercury had been different, then the atmospheric pressure would have to have been different, and so there would have been a different amount of rainfall, in accordance with (3). Lewis (1979) calls this kind of counterfactual a backtracking counterfactual. (3) is invariant under this sort of counterfactual change, so we have not yet found the right notion of invariance.

In addition to backtracking counterfactuals, Lewis also recognizes non-backtracking counterfactuals, and it is upon these that he erects his well-known analysis of causation (Lewis 1973, 2000). According to Lewis, the counterfactual 'if  $\underline{A}$  were true, then  $\underline{B}$  would be true', symbolized  $\underline{A} > \underline{B}$ , is true if in the closest possible worlds where  $\underline{A}$  is true,  $\underline{B}$  is true as well. Possible worlds are close to the actual world to the extent that they hold fixed the laws of nature and the particular matters of local fact that obtain in our world: the criteria for weighing the costs of various kinds of departures from actuality are given in Lewis (1979).

As we shall understand the notion of invariance, invariant generalizations must exhibit a pattern of non-backtracking counterfactual dependence. In particular, (3) will be invariant under changes in mercury level just in case some non-backtracking counterfactuals of the form: ‘if the height of the column of mercury ( $X$ ) had been  $x$ , then the amount of rainfall ( $Y$ ) would have been  $f(x)$ ’ are true. In fact, such claims are false: (3) is not invariant in this way and hence cannot successfully answer what-if-things-had-been-different questions and cannot serve in explanations. By contrast, Coulomb’s law does exhibit such a pattern of counterfactual dependence: it correctly answers questions about what would have happened had the charge density or geometry of the wire been different.

If Lewis’s account of non-backtracking counterfactuals were fully satisfactory, we could use it to provide an adequate account of the notion of invariance relevant to explanation and no further discussion would be needed. Although we think that Lewis’s account is more nearly correct than many of the competing treatments advanced by philosophers, we do not think that it is fully adequate. Our scepticism has several sources. First, Lewis’s criteria are vague in ways that causal claims do not seem to be. For example, those criteria require that we attach more importance to avoiding “big, widespread, diverse violations of law” than to avoiding “small, localized, simple violations of law” (1979 [1986], 48). We doubt that there is anything in scientific practice that tells us how to count miracles or to make such comparisons in a non-arbitrary way. Second, and more fundamentally, we doubt that it is possible to define criteria of similarity of worlds adequate for a theory of causal or non-backtracking counterfactuals in purely acausal terms, as Lewis attempts to do. That is, in order to determine whether (and how) the values of  $Y$  depend counterfactually on the values of  $X$ , one must make reference to the causal influence of other variables in specifying what must be held fixed. This view (or rather a closely analogous view) is now generally accepted among proponents of probabilistic theories of causation (e.g. Cartwright 1979, Eells 1991) and is gaining currency among writers on counterfactuals (e.g. Kvart 1986, Horwich 1987, Jackson 1977, and especially Pearl 2000, chapter 7). We will not defend our scepticism here, but will proceed directly to our positive theory.<sup>4</sup>

Central to our account of non-backtracking counterfactuals is the notion of an intervention. An intervention is an exogenous causal process that brings about the antecedent of the counterfactual in question. Heuristically, we may think of interventions as manipulations that might be carried out by a human being in an idealized experiment. Thus, in the case of (1), the possible worlds that are relevant to the evaluation of counterfactuals in which we imagine the charge density to be different are those in which we physically intervene to change the charge density along the wire by connecting it to an appropriate source or sink. Coulomb’s law is explanatory because it correctly tells us how the field intensity would change under such hypothetical interventions. A corresponding claim is not true of the barometer reading and the quantity of rainfall. Fiddling with a mercury column is not a way of bringing about or suppressing a storm, and this is why the former does not explain the latter.

These remarks about the connection between explanation and the results of human manipulation are intended only heuristically. It is not part of the theory we are proposing that causal and explanatory dependencies hold only when human intervention is possible. Nonetheless, we believe the heuristic value to be genuine. In particular, we take it to be an advantage of our approach that it makes clear the connection between counterfactuals, and the sorts of manipulations actually carried out in experiments used to test causal and explanatory claims. This connection is considerably more obscure on Lewis’s account.

To spell out more generally the requirements that an intervention must satisfy, let us begin with a simple experimental paradigm. A researcher wants to know whether treatment with a particular drug,  $D$ , causes recovery,  $R$ , from a particular disease. She divides a population of subjects, all of whom have the disease, into treatment and control groups. She administers the drug to all of the subjects in the treatment group, but to none in the control group, and then measures the frequency of recovery in the two groups. Her interventions  $I$  thus consist in some process that assigns each subject to one of these two groups, administering the drug to those in the treatment group and withholding it from those in the control group. She wishes to design the experiment in such a way

that any difference in recovery rates between the two groups can be attributed to the effect of the drug. What conditions must her experiment meet in order for this to be the case?

First, her interventions must determine whether or not any given subject receives the drug. This condition could be violated in a number of ways. Some members of the control group may already have quantities of the drug in their bloodstream, so that the experimenter's interventions do not succeed in withholding the drug from those subjects. Or subjects in the treatment group may fail to comply with the experimental protocol and throw their drugs away.

Second, the experimenter's interventions I must not be correlated with any factor that affects recovery, with the exception of those that lie along the causal chain from I to D to R. This condition could be violated, for example, if the subjects in the treatment group happen to have compromised immune systems, while those in the control group do not. Randomized trials are designed to avoid this sort of problem.

Our notion of intervention is a generalization of these restrictions. We will proceed in several stages. First, we shall introduce some concepts (such as the notion of one variable's being causally relevant to another) that will be needed in our definition. Second, we will use these concepts to define the notion of an intervention variable. Third, we will define the notion of an intervention, or more precisely, of a counterfactual whose antecedent is made true by an intervention. Fourth, we will define a special kind of intervention, which we call a testing intervention. Finally, we will use the notion of a testing intervention to formulate our account of explanation.

In the above informal discussion we have followed standard philosophical usage in treating causation as a relation between events or event-types, such as treatment with the drug (D) and recovery (R). For any given subject, each of these events occurs, or does not. However, in providing a more precise characterization, it is more perspicuous to follow the usual convention for the representation of causal relationships in the natural and social sciences and to treat variables as the primary causal relata. It is relatively easy to translate claims about events into claims about variables. In the above example, we may let the variable X take the value 1 or 0 according to whether some subject does or does not take the drug and let the variable Y take the value 1 or 0 according to whether or not the subject recovers. In general, we will say that X is causally relevant to Y if there is some set of circumstances<sup>5</sup> W in which the value of Y depends upon the value of X, i.e., there is some change in the value of X in W that would change the value Y or the probability that Y takes on some value. Thus in the above example, D is causally relevant to R if D either causes or prevents R in some subjects.<sup>6</sup> We emphasize that this notion of causal relevance is very permissive. In particular, X will count as causally relevant to Y if X affects the relationship between Z and Y, where Z is some other variable causally relevant to Y.

Based on this one example, it might seem that there is little to choose between the two modes of representation, event causation, and causal relevance between variables. This is indeed the case if we restrict attention to dichotomous variables such as X and Y. It is important to appreciate, however, that this is a special case. Consider the relationship between A, the amount of the drug given to a subject as measured by some continuous variable like mass, and T the amount of time that passes before some subject is healthy again. Suppose that up to some value larger drug doses will decrease recovery time, but that beyond this value they will harm the subject, delaying recovery. Does treatment with the drug 'cause' or 'prevent' speedy recovery? Armed with the notion of causal relevance between variables we may sidestep this issue: A is causally relevant to T and we may leave it at that. If we wish to be more informative, we should specify the form of the functional relationship between A and T. (See Hitchcock 1993 for further discussion.)

In addition to the basic notion of a variable X being causally relevant to a variable Y, we will need to make use of the notion of X's being causally relevant to Y via a route that excludes Z. This concept can be illustrated by means of a well-known example due to Hesslow (1976). One of the most worrisome possible side effects of birth control pills is the formation of blood clots: the consumption of birth control pills causes blood clots. On the other hand, birth control pills are very effective in preventing pregnancy, which can itself cause blood clots. If the latter effect is strong enough, it may well be that a woman is overall less likely to suffer from blood clots if she consumes birth control pills. In what sense, then, is it true that birth control pills cause blood clots? The natural answer is that birth control pills cause blood clots via a route that does not

include pregnancy (perhaps by the direct introduction of certain chemicals into the blood stream). If we employ the broadly manipulationist account of causation adopted in this essay, then we may also provide an intuitive elucidation of this notion in the following way: X is causally relevant to Y via a route that excludes Z if and only if there is some value of Z such that if we were to hold Z fixed at that value by means of an ideal experimental manipulation, and we were also carry out an ideal experimental manipulation that changes the value of X, then the value of Y (or the probability of Y assuming some value) would also change. For example, if a woman were to have a zygote implanted in her uterus, or to use some effective means of contraception other than birth control pills, this might well settle the issue of whether or not she becomes pregnant in a way that makes the consumption of birth control pills irrelevant to pregnancy. If birth control pills are causally relevant to blood clots via a route that excludes pregnancy, then the consumption of birth control pills would make a difference for the probability of blood clots even in one of these hypothetical situations where the woman's intervention fixes the value of the pregnancy variable<sup>7</sup>.

The final key concept that we will need is that of one variable acting as a switch for others. Suppose that a stereo receiver has three dials and one button on it: it has dials for volume, treble, and bass, as well as a power button. The setting of each of these is causally relevant to the frequency and amplitude of the soundwaves that are emitted from a speaker. Nonetheless, there is a certain asymmetry in the way in which these variables are relevant to the sound emitted by the speakers: the settings of the dials make a difference to the sound emitted only when the power button is in the 'on' position. When the power button is set to the 'off' position, there are no possible changes in the position of the other dials which will change the output of the speaker. In this case, the position of the power button is a switch for the settings of the other dials with respect to the output of the speaker. More generally a variable S acts as a switch for X with respect to Y if and only if there some is some value of S for which changes in X will change Y and some other value of S for which changes in X will not change Y. Informally, the value of the switch variable "interacts" with the value of X in such a way that when the switch is in the off position, the causal connection between X and Y is broken. We use this notion below to capture the idea that when an intervention on X occurs, the value of X (e. g., level of drug in the bloodstream) is entirely determined by the intervention; the previously existing endogenous causal connections (e. g. voluntary decisions by subjects about whether or not to take the drug) that have determined the value of X in the past no longer do so. The notion of a switch captures some of the features of the 'arrow-breaking' conception of interventions advocated by writers like Pearl (2000) and Spirtes, Glymour and Scheines (1993), while avoiding other features of this idea that some have found objectionable.

In order to define the key notion of an intervention, we must first define what it is for a variable I to be an intervention variable for X, with respect to Y.

Let X be a variable, whose values represent various properties that might be possessed by some individual, and let Y be another variable (not necessarily applying to the same individual). Then I is an intervention variable for X, with respect to Y, if it meets the following conditions:<sup>8</sup>

- (M) 1) I is causally relevant to X.
- 2) I is not causally relevant to Y through a route that excludes X.<sup>9</sup>
- 3) I is not correlated with any variable Z that is causally relevant to Y through a route that excludes X, be the correlation due to I's being causally relevant to Z, Z's being causally relevant to I, I and Z sharing a common cause, or some other reason.
- 4) I acts as a switch for other variables that are causally relevant to X. That is, certain values of I are such that when I attains those values, X ceases to depend upon the values of other variables that are causally relevant to X.

Returning to our idealized experiment, I would be a variable taking as possible values {assign to treatment group, assign to control group, do not intervene}, X would represent taking the drug or not, and Y would represent recovery. Clauses 1 and 4 require that the value of I make a difference

for whether or not a subject takes the drug, and also that when a subject is assigned to one of the two groups, that  $I$  be the only variable whose value makes a difference to whether or not a subject takes the drug. Clauses 2 and 3 require that the only effect of  $I$  on  $Y$  be through its effect on  $X$ .

If we are willing to make some additional assumptions, it is possible to simplify (M) considerably. Suppose that we adopt a framework similar to that of Spirtes, Glymour and Scheines (1993), and assume that there is a well defined probability distribution over the values of the variables. Assume also that this distribution satisfies the so-called Causal Markov condition, which says that conditional on its direct causes, each variable is independent of every other variable except its effects. (This is a generalization of the familiar screening-off assumptions frequently adopted in discussions of probabilistic causality — joint effects are screened off from one another by the full set of their common causes, distal causes are screened off from their effects by direct or proximal causes, and so on.)<sup>10</sup> Then clauses 2 and 3 reduce to the following:  $I$  is probabilistically independent of  $Y$ , conditional on the value of  $X$ . We retain the more general formulation in (M), in part to presuppose as little as possible about the nature of causation and its relation to probability, and in part to present an explicit list of what is excluded by the concept of an intervention variable.

An intervention on  $X$  with respect to  $Y$  is an actual or hypothetical change in the value of some variable  $I$ , where  $I$  is an intervention variable for  $X$  with respect to  $Y$ . Armed with the concept of an intervention, we can now state our account of non-backtracking counterfactuals. A counterfactual of the form 'if  $X$  were to have the value  $x$ , then  $Y$  would have the value  $y$ ' is true if and only if  $Y$  has the value  $y$  in the hypothetical situations (or possible worlds) where (i) the value of  $X$  is equal to  $x$ ; and (ii) all other variables have their actual values with the exception of  $I$ , and any variable for which  $I$  is causally relevant, where  $I$  is an intervention variable for  $X$  with respect to  $Y$ . More intuitively, we are to imagine a situation in which the value of  $X$  is changed to  $x$  as a result of the change in the value of a variable that meets condition (M). Such an intervention is 'surgical' or 'minimal' in the sense that only the values of variables that are effects of the intervention are changed. As advertised, our notion of an intervention makes no essential reference to human beings or their activities; instead it is characterized purely in terms of notions like causal relevance and probabilistic independence. Thus a purely natural process not involving human activity at any point may qualify as an intervention as long as it satisfies the conditions described by (M).

We emphasize that the conditions (M) are stated in terms of concepts (such as causal relevance) that are overtly causal. This means that one cannot appeal to the notion of an intervention as part of a reductive account of what it is for  $X$  to be causally or explanatorily relevant to  $Y$ . However, we should also note that our account of what qualifies  $I$  as an intervention variable for  $X$  with respect to  $Y$  makes no reference to the presence or absence of a causal relationship between  $X$  and  $Y$ . Instead (M) makes reference to other causal relationships: the causal relationship between  $I$  and  $X$ , the causal and probabilistic relationships between  $I$  and various other causes of  $Y$  besides  $X$ , and so on. In particular, we should note that requiring that  $I$  affect  $Y$  if at all only through  $X$  is not tantamount to requiring that  $I$  does affect  $Y$  through  $X$ . Moreover, our account of intervention makes reference only to qualitative causal concepts, and not to the quantitative or functional form of the relationship between variables. Thus, although non-reductive, our account is not viciously circular: it does not presuppose the very thing that it aims to assess — whether a particular functional relationship  $Y = f(X)$  is invariant under interventions.

Some philosophers hold that non-reductive accounts of causation and explanation must inevitably be unhelpful and unilluminating. We think, in agreement with a growing number of writers both within and outside of philosophy, that this attitude is fundamentally mistaken. The conclusion that it is impossible to analyze causal relationships in terms of acausal concepts like 'correlation' is now widely accepted both among philosophers working on probabilistic theories of causation and by thoughtful statisticians, econometricians and theorists of experimental design. There is general agreement among these researchers that adoption of a non-reductive account of causation does not preclude clear and precise treatments of causal claims themselves or of the epistemological problems surrounding inference to such claims. Moreover, analogous conclusions have been reached elsewhere in philosophy. For example, there is general agreement

that it is impossible to characterize what it is for a subject  $S$  to believe that  $p$  just in terms of 'non-mentalistic' notions that refer only to overt behavior. However, it is arguable that it is possible to characterize what it is for  $S$  to believe that  $p$  if one is allowed to make reference to  $S$ 's other beliefs and desires, as well as to her overt behavior. Such a characterization is non-reductive, but it is not viciously circular. So also for the account we propose.

A closely related point is that while our account presupposes some causal notions, it embodies a number of non-trivial substantive assumptions about the interrelationships among other concepts ('intervention', 'counterfactual dependence', 'explanation') in the same family. These assumptions are in turn inconsistent with a number of received views about explanation. For example, we will argue below that if the interventionist based account that we advocate is correct, it follows that standard accounts of lawfulness and explanation focus on the wrong sort of counterfactuals. The correctness of this claim is quite independent of the issue of whether causal and explanatory notions can be given a reductive characterization. More generally, our account illustrates the point that a theory can be non-reductive without being trivial or uninformative.

Finally, a word about 'miracles'. Consider a counterfactual of the form 'if  $X$  had taken the value  $x$ , then  $Y$  would have taken the value  $y$ ', where it is understood that the  $X$  is changed to  $x$  by means of an intervention variable  $I$  taking the value  $i$ . The notion of an intervention has been designed in such a way that, as long as  $I$  is an intervention variable, it makes no difference to the value of  $Y$  how  $I$  comes to possess the value  $i$ . We can, if we wish, say that  $I$  acquires this value as a result of some minor miracle. Alternatively, we can imagine that  $I$  is caused to have the value  $i$  as a result of a change in the value of some other variable  $Z$  which is causally relevant to  $I$ , that the value of  $Z$  changes because of a change in the value of yet another variable  $W$  and so on, with any needed miracle occurring at some much earlier time. Because it is built into the notion of an intervention that any change in the value of  $Y$  will occur only through the change in the value of  $X$ , the details of how  $I$  comes to have the value  $i$  do not matter to the assessment of the above counterfactual. In other words, the use of an intervention variable allows us, if we wish, to push any required miracle back indefinitely, so that the laws of the actual world continue to hold in the region of interest. This provides a natural explanation of a striking feature of the kinds of counterfactuals that are relevant to causal and explanatory claims: that while we require that they be true when their antecedents are realized by interventions, any more detailed specification of the way in which their antecedents are realized is regarded as irrelevant and unnecessary.

We are now in a position to characterize the notion of invariance: a relationship  $R$  between variables  $X$  and  $Y$  is invariant if it would continue to be true (or approximately true) in at least some hypothetical situations or possible worlds in which the value of  $X$  is changed as the result of an intervention. That is, there must be some non-actual value  $x$  of  $X$  such that the following counterfactual is true: 'if  $X$  were equal to  $x$ , then the values of  $X$  and  $Y$  would stand (approximately) in the relationship  $R$ '. This account can be naturally extended to cover cases where  $X$  and  $Y$  are sets of variables, rather than individual variables.

We emphasize that our account of invariance under interventions is existential, rather than universal in character: to count as invariant an explanatory generalization must be invariant under some — not necessarily all — interventions. As noted earlier, many and perhaps most explanatory generalizations hold only for some range of interventions and break down under others. When there are no interventions with respect to some generalization (i. e. when no interventions are 'possible'), our account has the consequence that the generalization fails to be invariant under interventions, rather than being trivially invariant.

What does it mean to say that there are no interventions with respect to some generalization? We do not require that interventions be physically possible in the sense of being consistent with physical law and actually obtaining initial conditions: an intervention may well require a miracle somewhere in its history. For example, in the case of Coulomb's law, we take it to be meaningful to talk about interventions to change the charge density along the wire, even if the actual charge density was determined to be what it was. This sort of case may be contrasted with cases in which the existence of a certain sort of intervention is precluded by physical law alone (rather than the laws and the actual initial conditions together). In at least some cases of this sort, it seems plausible that there exist no interventions for some generalization of interest to be invariant under

and hence that the generalization is non-explanatory. For example, in EPR type set-ups, as a matter of physical law alone there is no intervention on the outcome of the left hand measurement, with respect to the outcome of the right hand measurement. That is, any method of bringing about a desired outcome on the left hand side (such as preparing the particle pairs in a state other than the singlet state) will have a direct effect on the outcome on the right hand side as well. It is thus natural to think of this as a case in which the perfect (anti)-correlation between the two measurement outcomes in an EPR experiment is not invariant under (any) interventions, and it follows that one measurement result does not cause or explain the other.<sup>11</sup>

In still other cases, while there may be no physical law forbidding interventions, interventions may be ill-defined in the sense that we lack any clear notion of what it would be to change a system in the way envisioned or of what would be true under such a change. For example, we doubt that there is any clear notion of an intervention that would change Bill Clinton into a copper wire or Adam Morton into a dry, well-made match.<sup>12</sup> This point will play a central role in the discussion of section 4 below.

There is a final restriction we wish to impose upon the kind of invariance under intervention that matters for successful explanation. Imagine a light bulb that is normally triggered by a circular switch. The switch has a little hash mark, that can be moved from a vertical position ( $0^\circ$ ) to just past horizontal ( $100^\circ$ ). At a threshold one radian (approximately  $57^\circ$ ), the light comes on. We can represent this relationship as follows. Let  $\underline{L}$  be a variable that takes the value 1 if the light is on, 0 otherwise; let  $\underline{\theta}$  be the angular displacement of the switch measured in radians; and let  $\lfloor \_ \rfloor$  be the whole part function, so that for example  $\lfloor 1.57079... \rfloor = 1$ . Then the relationship between  $\underline{L}$  and  $\underline{\theta}$  is:

$$(4) \quad \underline{L} = \lfloor \underline{\theta} \rfloor.$$

(Note that the maximum angular displacement of the switch was stipulated to be  $100^\circ$ , which is less than 2 radians.) This relationship is invariant under interventions on  $\underline{\theta}$ : if we intervene to change the position of the switch, say from 0 to  $\pi/2$  ( $90^\circ$ ), the light will go from off to on. It is intuitively plausible, that if the switch is turned to  $90^\circ$  and the light is on, the relationship (4) can figure in an explanation of why the light is on. So far, so good.

Suppose that the switch is turned to an angle of  $90^\circ$  and the light is on, as before. However, now we imagine the light switch to be broken, so that the light will remain on no matter what the setting of the switch. In this case, it seems wrong to say that relationship (4) figures in an explanation of the light's being on. The light's being on is completely independent of the position of the switch. Nonetheless, it is still the case that (4) is invariant under some interventions on  $\underline{\theta}$ : we can intervene to set  $\underline{\theta}$  anywhere from 1 to  $\pi/2$  radians and (4) will continue to hold.

In order to deal with this sort of case, we propose a further restriction on the interventions that are to count for purposes of determining the explanatory credentials of some relationship. Let  $\underline{R}$  be a relationship that holds between the actual values of certain variables, and suppose that  $\underline{R}$  figures in a putative explanation of why the explanandum variable has the value it does. We will say that  $\underline{R}$  is invariant under testing interventions if it is invariant under interventions that change the values of the other variables in such a way that  $\underline{R}$  predicts a value for the explanandum variable different from the value it actually had. For example, if the light is on and the switch is turned to an angle of  $90^\circ$ , a testing intervention for relation (4) would be one that sets  $\underline{\theta}$  to less than one radian, where (4) predicts a change in the value of  $\underline{L}$ . Then we should require that explanatory relationships be invariant under testing interventions. In the case where the switch mechanism is broken, (4) fails to be invariant under testing interventions, since setting  $\underline{\theta}$  to less than one radian does not put out the light. An explanatory generalization must tell us about how different values of the explanandum variable would result from interventions that change the values of the explanans variables. It is only if a generalization meets this condition that it can figure in answers to what-if-

things-had-been-different questions. Thus in the informal statement in section 2, the requirement that the explanans of a successful explanation include only invariant relationships should be understood in the following sense: these relationships must be invariant under testing interventions on the values of the explanans variables. In what follows, we will continue to use the shorthand ‘invariant’ for ‘invariant under testing interventions on the values of the explanans variables’.

#### 4. The Nomothetic Conception Revisited

With a clearer understanding of the key concepts of invariance and intervention, we may contrast our account with Hempel’s D-N model in greater detail, with particular emphasis on how invariant generalizations differ from laws of nature.

Laws have traditionally been understood as universal generalizations of the form ‘All A’s are B’s’. It is our contention that such generalizations, even when they satisfy the conditions for lawhood standardly imposed by philosophers, often fail to be explanatory, or are at best very poor explainers. Discussions of the D-N model of explanation have often employed toy examples of explanations having the logical form:

- (5)            All A’s are B’s  
                   Object o is an A  
                   Therefore, o is a B

These have always had an air of artificiality about them. Real scientific explanations are much more complex affairs with considerable additional structure. Philosophers of science have generally recognized this but have nonetheless assumed, no doubt under the influence of the D-N model, that (5) is an acceptable idealization — that it captures all the essential features of genuine explanations.

Note that (5) does not have the appropriate form to be an explanation on our account, since we require that the explanans and explanandum be expressed in terms of the values of variables. This is easily remedied, however. Let X be the characteristic function or indicator function for A, so that for any object o,  $X(o) = 1$  if and only if o is A, and  $X(o) = 0$  if and only if o is not A. Analogously, let Y be the characteristic function for B. Then (5) can be re-written:

- (5’)           For all objects,  $X \leq Y$   
                   For object o,  $X(o) = 1$   
                   Therefore,  $Y(o) = 1$

So the mere fact that (5) does not explicitly specify relationships between variables is no objection to its qualifying as an explanation.

The problem, rather, is that the generalizations figuring in (5) and (5’) do not tell us what being a B (or having a certain value of Y) depends on. It may be that all objects are B’s, or at any rate that all objects belonging to a much broader class are B’s. Consider, for example, the generalization:

- (6)            All igneous rocks have mass greater than zero.

While this fits nicely with the traditional conception of a law, it does nothing to tell us why some particular rock is massive: its having a non-zero mass in no way depends upon its petrologic classification. This notion of dependence is captured by our notion of a testing intervention: it is only if a generalization is invariant under testing interventions that it conveys information about how one variable depends on another. By subjecting an igneous rock to tremendous heat and pressure, it can be transformed into a metamorphic rock. This is an intervention, but not a testing intervention for the generalization in question: metamorphic rocks are also massive. Nor are there any other interventions that qualify as testing interventions under which this generalization is invariant — hence our sense that it tells us nothing about what the massiveness of rocks depends on. Put another way, the generalization cannot be used to answer what-if-things-had-been-different

questions about hypothetical situations in which the rock would not have been massive. Indeed, if anything, (6) suggests false answers to questions about what would happen if a particular rock were transformed from igneous to metamorphic.

This failure is closely connected to the familiar problem of explanatory irrelevance. The classic example comes from Salmon (1971):

- (7) All men who regularly consume birth control pills fail to become pregnant.  
John Jones regularly consumed birth control pills.  
Therefore, John Jones failed to become pregnant.

The generalization in (7), while arguably a law, is not invariant under testing interventions. In particular, intervening to prevent John Jones from taking birth control pills would not affect his chances of conception. Thus, while (7) satisfies the requirements for successful D-N explanation, it does not identify conditions such that changes in those conditions resulting from interventions would lead to changes in the explanandum. According to our account, this is why it fails to be explanatory.<sup>13</sup>

Our account agrees with Hempel's that to be explanatory with respect to an object o, a generalization relating X and Y must not merely say that the actual values of X and Y possessed by object o stand in this relation. And like Hempel, we agree that the additional content of explanatory generalizations is (at least partially) counterfactual in nature. However, our account disagrees with Hempel's (and with other standard treatments of the role of laws in explanation) about what this additional content consists in. According to the standard view, to count as a law, and hence as explanatory, a generalization must describe a relationship that holds of the actual values of X and Y possessed by objects other than o. Moreover, the relationship must also hold for certain hypothetical values possess by objects other than o. In particular, if a generalization like 'All A's are B's' is to be explanatory with respect to o, it must 'support' counterfactuals of the following form: if some object o\* that is different from o and does not possess property A were to be an A, then it would be a B. (In terms of our schema (5'), if o\*, for which  $X(o^*) = 0$ , were such that  $X(o^*) = 1$ , then  $Y(o^*) = 1$ .) We will refer to counterfactuals of this sort as 'other object' counterfactuals.

By contrast, according to our account, to count as invariant and hence explanatory, a generalization must describe a relationship that holds for certain hypothetical values of X and Y possessed by the very object o, namely those values of X and Y possessed by o in the hypothetical situations where the value of X is changed by an intervention. That is, on our account, the counterfactuals that must be supported by a generalization that is explanatory with respect to o are of the following form: if the value assigned by the variable X to o were to be changed via an intervention (e g., from  $X(o) = 1$  to  $X(o) = 0$ ), then the value assigned by Y to o would change in some way predicted by the generalization. Let us call these 'intervention' counterfactuals.

We maintain that it is the ability of a generalization to support such intervention counterfactuals, rather than its ability to support other object counterfactuals, that determines its explanatory potential. Consider the example in which Coulomb's law is used to explain why a particular conductor with such and such a geometrical configuration and charge distribution produces an electric field potential of such and such a strength. According to the traditional view, Coulomb's law is considered a genuine law (and hence as explanatory), because it supports other object counterfactuals, including (apparently) such counterfactuals as: 'if Bill Clinton, the H.M.S. Victory, or a neutron were a conductor with such and such a geometrical configuration and charge distribution, he/she/it would produce an electric field potential of such and such a strength.' Such counterfactuals tell us very little about what the actual field potential depends on, in part because it is so hard to comprehend their antecedents, and in part because they involve changes in the identity of the conductor, and this is not a factor on which the strength of the field depends. Instead the strength of the field produced by a conductor depends on such factors as its geometry and charge density.

On our view Coulomb's law is explanatory because it supports counterfactuals about what would happen if the charge density or geometric configuration of this very conductor were changed in various ways as a result of an intervention. This counterfactual information does make explicit how the strength of the potential field produced by the conductor depends upon its geometry and charge distribution. In contrast to 'other object' counterfactuals, the counterfactuals on which our account focuses exhibit explanatorily relevant information.

It is ironic that the traditional conception of laws is often defended on broadly empiricist grounds: universality is said to be an empirically respectable notion, while richer modal notions are not. It is wholly mysterious how we might test counterfactuals about what would happen if Bill Clinton, the H.M.S. Victory, or a neutron were to be a long, straight wire; indeed, it is mysterious what such counterfactuals mean. By contrast, the counterfactuals that are central to our account are often directly empirically testable. For instance, it may be possible to intervene to change the geometry or charge distribution of the conductor in order to determine whether the relationship expressed by Coulomb's law continues to hold for this conductor. We think that the difference between these two sorts of counterfactuals is reflected in the judgments of most scientists about the physical content of Coulomb's law. While it is natural and intuitive to think of that law as telling us how the field due to some conductor would change as the charge density or geometry of a conductor were changed, the law is simply not in the business of trying to tell us what would happen under the fantastic transformations contemplated above.

## 5. Conclusion

We have argued that to explain why some phenomenon occurs is to show what that phenomenon depends upon. This is achieved by providing the resources for answering a variety of what-if-things-had-been-different questions: how would the outcome have differed if the initial conditions had been changed in various ways? A generalization can play this role if and only if it is invariant under interventions. Such a generalization supports a range of counterfactuals: not the 'other object' counterfactuals upon which traditional discussions of laws have focussed, but counterfactuals about what would happen under certain interventions on the system at hand.

This account of explanation has a number of virtues. It allows us to diagnose standard counterexamples to the nomothetic conception of explanation that involve failures of explanatory relevance. Such examples, while they conform to the structure of D-N arguments, do not provide us with the resources for answering what-if-things-had-been-different questions. Moreover, our account allows us to avoid a standard dilemma involving explanation in the social sciences: either the generalizations of the special sciences are (despite all appearances to the contrary) 'laws' or else they are unexplanatory. On our account, the generalizations of the special sciences can be used to explain if they are invariant under interventions, regardless of whether they qualify as 'laws' in the traditional sense. Finally, our account is able to make sense of the very natural intuition that some explanations are deeper and more powerful than others. We will argue for this claim in detail in a forthcoming companion piece (Hitchcock and Woodward forthcoming).

## REFERENCES

- Beatty, J. (1995) "The Evolutionary Contingency Thesis," in Concepts, Theories and Rationality in the Biological Sciences, ed. J. Lennox and G. Wolters, Pittsburgh: University of Pittsburgh Press, pp. 45 -81.
- Butterfield, J. (1992) "David Lewis Meets John Bell," Philosophy of Science 59: 26 - 43.
- Cartwright, N. (1979) "Causal Laws and Effective Strategies," Noûs 13: 419-437.
- Cartwright, N. and M. Jones (1991) "How To Hunt Quantum Causes," Erkenntnis 35: 205-31.
- Earman, J., and J. Roberts (1999) "Ceteris Paribus, There is No Problem of Provisos," Synthese 118: 439 - 478.
- Eells, E. (1991) Probabilistic Causality, Cambridge: Cambridge University Press.
- Fay, B. (1983) "General Laws and Explaining Human Behavior," in Changing Social Science, ed. D. Sabia and J. Wallulis, Albany: SUNY Press, pp. 103 - 128.
- Freedman, D. (1997) "From Association to Causation Via Regression," in Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences, ed. V. McKim and S. Turner, Notre Dame: University of Notre Dame Press, pp. 113 - 161.
- Halpern, J. and J. Pearl (2000) "Causes and Explanations: A Structural-model Approach," Technical report R-266, Cognitive Systems Laboratory, University of California at Los Angeles.
- Hausman, D. (1998) Causal Asymmetries, Cambridge: Cambridge University Press.
- Hausman, D., and J. Woodward, (1999) "Independence, Invariance and the Causal Markov Condition," British Journal for the Philosophy of Science 50.
- Hempel, C. (1965a) "Aspects of Scientific Explanation," in Hempel 1965b, pp. 331 - 496.
- Hempel, C. (1965b) Aspects of Scientific Explanation and Other Essays in the Philosophy of Science, New York: Free Press.
- Hempel, C., and P. Oppenheim (1948) "Studies in the Logic of Explanation," Philosophy of Science 15: 135 - 175. Reprinted in Hempel 1965b, pp. 245-290.
- Hesslow, G. (1976) "Discussion: Two Notes on the Probabilistic Approach to Causality," Philosophy of Science 43: 290 - 292.
- Hitchcock, C. (1993) "A Generalized Probabilistic Theory of Causal Relevance," Synthese 97: 335 - 64.
- Hitchcock, C. (1995) "Discussion: Salmon on Explanatory Relevance," Philosophy of Science 62: 304 - 320.
- Hitchcock, C. (2001). "The Intransitivity of Causation Revealed in Equations and Graphs." Journal of Philosophy 98, 6: 1 - 27.
- Hitchcock, C. (Forthcoming). "A Tale of Two Effects."
- Hitchcock, C., and J. Woodward (Forthcoming) "Explanatory Generalizations, Part II: Plumbing Explanatory Depth," Noûs.
- Horwich, P. (1987) Asymmetries in Time, Cambridge, MA: MIT Press.
- Jackson, F. (1977) "A Causal Theory of Counterfactuals," Australasian Journal of Philosophy 55: 3 - 21.
- Kvart, I. (1986) A Theory of Counterfactuals, Indianapolis: Hackett.
- Lewis, D. (1973) "Causation," Journal of Philosophy 70: 556 - 567. Reprinted with Postscripts in Lewis 1986, pp. 159 - 213.
- Lewis, D. (1979) "Counterfactual Dependence and Time's Arrow," Noûs 13: 455 - 476. Reprinted with Postscripts in Lewis 1986, pp. 32 - 66.
- Lewis, D. (1986) Philosophical Papers, Volume II, Oxford: Oxford University Press.
- Lewis, D. (2000) "Causation as Influence," Journal of Philosophy 97: 182 - 197.
- Mackie, J. (1974) The Cement of the Universe, Oxford: Oxford University Press.
- Meek, C. and C. Glymour (1994) "Conditioning and Intervening," British Journal for the Philosophy of Science 45: 1001 -21.
- Morton, A. (1973) "If I Were a Dry Well-Made Match," Dialogue 12: 322-324
- Pearl, J. (2000) Causality, Cambridge: Cambridge University Press.
- Pietroski, P. and G. Rey (1995) "When Other Things Aren't Equal: Saving Ceteris Paribus Laws from Vacuity," British Journal for the Philosophy of Science 46: 81 - 110.

- Rosenberg, A. (1992) Economics: Mathematical Politics or Science of Diminishing Returns, Chicago: University of Chicago Press.
- Salmon, W. (1971) "Statistical Explanation," in Statistical Explanation and Statistical Relevance, ed. W. Salmon, Pittsburgh: University of Pittsburgh Press, pp. 29 - 87.
- Salmon, W. (1984) Scientific Explanation and the Causal Structure of the World, Princeton: Princeton University Press.
- Salmon, W. (1989) Four Decades of Scientific Explanation, Minneapolis: University of Minnesota Press.
- Scriven, M. (1956) "A Possible Distinction between Traditional Scientific Disciplines and the Science of Human Behavior," in Minnesota Studies in the Philosophy of Science, Volume I, ed. H. Feigl and M. Scriven, Minneapolis: University of Minnesota Press, pp. 330 - 339.
- Skyrms, B. (1980) Causal Necessity, New Haven: Yale University Press.
- Smart, J. J. C. (1963) Philosophy and Scientific Realism, London: Routledge and Kegan Paul.
- Spirtes, P., C. Glymour, and R. Scheines (1993) Causation, Prediction, and Search, New York: Springer-Verlag.
- Woodward, J. (1979) "Scientific Explanation," British Journal for the Philosophy of Science 30: 41 - 67.
- Woodward, J. (1984) "A Theory of Singular Causal Explanation," Erkenntnis 21: 231 - 262.  
Reprinted in Explanation, ed. D. Ruben. Oxford: Oxford University Press, 1993, pp. 246-274.
- Woodward, J. (1997a) "Explanation, Invariance and Intervention," PSA 1996, vol. 2, S26- 41.
- Woodward, J. (1997b) "Causal Modeling, Probabilities and Invariance," in Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences, ed. V. McKim and S. Turner. Notre Dame: University of Notre Dame Press, pp. 265 - 317.
- Woodward, J. (1999) "Causal Interpretation in Systems of Equations," Synthese 121: 199-257.
- Woodward, J. (2000) "Explanation and Invariance in the Social Sciences," British Journal for the Philosophy of Science 51: 197-254.
- Woodward, J. (Forthcoming) "Causality and Manipulation".

## NOTES

\* This paper had its origins in a talk given by Woodward entitled “Explanation and Invariance” at the 1997 Eastern Division Meetings of the American Philosophical Association and the commentary on the talk by Hitchcock. Marc Lange was the other commentator and we are grateful to him for a number of helpful comments and suggestions. We are also grateful to Nancy Cartwright, Malcolm Forster, Alan Hájek, Dan Hausman, Paul Humphreys, and Judea Pearl for helpful discussions. Woodward’s contribution to this paper was supported in part by the National Science Foundation (SBR-9320097).

<sup>1</sup> For a more detailed discussion of this idea and of the causal claims implicit in equations like (2), see Woodward (1999) and Hausman and Woodward (1999).

<sup>2</sup> Discretion being the better part of valor, we will remain neutral on the issue of whether it is absorption or transmission *per se*, or merely the probability of absorption and transmission that is explained on any particular occasion.

<sup>3</sup> How to extend a ‘basic’ account of causation to cover more complex cases is a problem faced by all theories of singular causation, and numerous proposals have been offered (see for example Lewis 1973, 2000). For more recent accounts that are very much in the spirit of the ideas developed in this essay, see Pearl (2000, chapter 10), Halpern and Pearl (2000), and Hitchcock (2001).

<sup>4</sup> From our perspective, Lewis’s theory is as successful as it is in reproducing our ordinary causal judgments because his notion of a minor miracle does roughly the same work as the notion of an intervention in our account. Nonetheless, the two approaches do not yield the same causal judgments in all cases — see note 10 below.

<sup>5</sup> The allowable set of possible circumstances must be constrained somehow. Any variable  $X$  will count as causally relevant to  $Y$  if we allow as circumstances ‘the existence of a powerful deity who sets the value of  $Y$  according to the value of  $X$ .’ We have in mind that in any given case of explanation there is some set of conditions that is being ‘held fixed’ in the background. In example (2), for instance, it might be understood that the plant is growing on the earth, and thus that circumstances in which there is a different gravitational field or atmosphere are ‘out of bounds’. Mackie (1974) refers to these background conditions as the ‘causal field’. In the causal modeling approach as developed by Spirtes et al. (1993) and Pearl (2000), assumptions about which sets of circumstances are considered relevant are reflected in the choice of variables to include in the model: see Hitchcock (2001) for more discussion.

<sup>6</sup> Note that our term ‘causally relevant’ is unlike the term ‘causally connected’ used by Hausman (1998): we do not say that  $X$  is causally relevant to  $Y$  if values of  $Y$  cause values of  $X$ , or if  $X$  and  $Y$  are effects of a common variable.

<sup>7</sup> Both Hitchcock (forthcoming) and Woodward (forthcoming) contain more detailed characterizations of the notion of a causal route.

<sup>8</sup> There are a number of other characterizations of the notion of an intervention in the recent philosophical and statistical literature. Most of these largely coincide with our characterization but sometimes differ in detail. See, e. g. Spirtes, Glymour and Scheines (1993), Cartwright and Jones (1991), Meek and Glymour (1994), Pearl (2000), Hausman (1998). We do not have the space to explore in detail the relationship between these various accounts.

<sup>9</sup> Note that because of our liberal conception of what it means for one variable to be causally relevant to another, M2 will exclude the sort of case described by Cartwright and Jones (1991) in which  $I$  affects the mechanism connecting  $X$  and  $Y$ .

<sup>10</sup> For a more precise statement of this condition, see Spirtes, Glymour and Scheines, (1993, 54). For further discussion, see Woodward (1997b) and Hausman and Woodward (1999).

<sup>11</sup> Contrast this with the apparent consequence of Lewis’s account that the outcome on one wing causes the outcome on the other (see Butterfield 1992). This is just one of a number of cases in which our account apparently yields different causal conclusions than Lewis’s.

<sup>12</sup> The allusion is to Morton’s wonderfully titled paper: “If I Were a Dry, Well-made Match” (Morton 1973).

<sup>13</sup> For further discussion, see Woodward (2000, section 3).