Data, Phenomena, and Reliability

Jim Woodward†‡
California Institute of Technology

This paper explores how data serve as evidence for phenomena. In contrast to standard philosophical models which invite us to think of evidential relationships as logical relationships, I argue that evidential relationships in the context of data-to-phenomena reasoning are empirical relationships that depend on holding the right sort of pattern of counterfactual dependence between the data and the conclusions investigators reach on the phenomena themselves.

1. Introduction. In a series of papers written some years ago (Bogen and Woodward 1988, 1992, forthcoming; Woodward 1989), Jim Bogen and I introduced a distinction between data and phenomena. Phenomena are stable, repeatable effects or processes that are potential objects of prediction and systematic explanation by general theories and which can serve as evidence for such theories. For example, neutral currents are a phenomenon which is explained by, and indeed served as a decisive piece of evidence for, the Weinberg-Salam electroweak theory. The deflection of starlight by the sun measured by the Eddington eclipse expedition of 1919 is a phenomenon which is predicted and explained by General Relativity. Data are public records (bubble chamber photographs in the case of neutral currents, photographs of stellar positions in the case of Eddington's expedition) produced by measurement and experiment, that serve as evidence for the existence of phenomena or for their possession of certain features. When data play this role they reflect the causal influence of the phenomena for which they are evidence but they also reflect the operation of local and idiosyncratic features of the measurement devices and exper-

†Send requests for reprints to the author, Division of the Humanities and Social Sciences, 101–40, California Institute of Technology, Pasadena, CA; e-mail: jfw@hss.caltech.edu.

‡Thanks to my co-symposiasts Peter Achinstein and Deborah Mayo, and also to Chris Hitchcock, Allan Franklin, and Elliott Sober for very helpful comments.

Philosophy of Science, 67 (Proceedings) pp. S163–S179. 0031-8248/2000/67supp-0014$0.00

imental designs that produce them. For example, Eddington's photographs are the net upshot of an extremely complex combination of causal factors that include not just the stellar light deflected by the sun's gravitational field but also such local features as the characteristics of Eddington's telescopes and tracking machinery, the effects of local variations in temperature on both the telescopes and the photographic plates, the chemical processes used to develop the plates, human decisions about the placement of equipment and so on.

Because data are the result of such complex and idiosyncratic processes involving many different sorts of causal factors it is often difficult to derive or systematically explain their detailed features from general theory in conjunction with what philosophers usually think of as background information about initial and background conditions. Thus General Relativity, in conjunction with information about initial conditions such as the mass distribution of the sun, the positions of the stars, etc., allows us to predict and explain the deflection of starlight but not detailed descriptions of Eddington's photographs. More importantly, and whether or not one thinks that such derivations must be possible "in principle" when General Relativity is supplemented with an appropriate theory of Eddington's measuring apparatus, such derivations often do not play an important role in the patterns of reasoning from data to phenomena that scientists actually employ. In particular the role of the photographs as evidence has little to do with our ability to construct (nontrivial) deductive or even inductive derivations of features of them from the fact of starlight deflection. Instead what matters is (roughly) whether we can use the photographs to discriminate reliably among different possible values for the starlight deflection. The direction of inference in such cases is "upwards" from the data to some feature of the phenomenon, rather than "downwards" from the phenomenon to features of the data. Typical inferences from data to phenomena thus represent one important example of a more general fact which has been discussed by Achinstein (1991, 308), namely, that evidence for a claim need not be (deductively or inductively) derived from it.

The data/phenomena framework contrasts with a more traditional picture according to which scientific theories contain deductive connections or rules that directly associate theoretical claims with so called observational claims. The framework Bogen and I advocate replaces this two-part structure with a more complex structure containing at least three components: theoretical claims, phenomena claims, and data, where the connection between the last two items is not prescribed by the theory that explains the phenomena claims but rather by independent reasoning which is often empirical in character (see Section 4). This more complex structure fits with another feature of scientific practice noted by Achinstein (1991,

314), namely, that general theories rarely provide instructions or guidance about how to construct experiments that will produce data to test the phenomena claims predicted by such theories.

2. Data Production and Data Interpretation. It is useful to distinguish two components in the overall process by which data are used to reach conclusions about phenomena. I will call them *data production* and *data interpretation*. Data production has to do with the causal processes that lead from the phenomenon of interest to the data. It involves setting up systems of causal interaction or locating preexisting systems which allow for the production of effects which can be perceived and interpreted by investigators and which permit investigators to discriminate among competing phenomena claims. Data interpretation involves the use of arguments, analytic techniques, and patterns of reasoning which operate on the data so produced to reach conclusions about phenomena. Typical components of this stage are techniques of data analysis and reduction, including statistical procedures of various sorts and procedures for smoothing, transforming, and discarding data. Data interpretation may also include the explicit use of background or theoretical assumptions, as when kinematic assumptions are used to calculate the momentum of some particle from a bubble chamber photograph or when premises about the decay rate and branching ratio of radioactive potassium are used to calculate the age of some geological formation. While changes in data production will involve changes in the physical processes that cause the data and very often changes in the physical characteristics of the data themselves, changes in data interpretation will leave these processes and characteristics unaltered and will instead involve changes in the assumptions and patterns of reasoning that investigators bring to the data after it has been produced. Thus, changing the focal length or lens of a telescope alters the process of data production and the data produced, while changing the statistical procedures used to assess the photographs it produces amounts to a change in data interpretation. Similarly, changing from an observational study in which data is produced non-experimentally to a design in which data are produced by a randomized experiment changes the data production process. When we change the econometric techniques that we use to analyze the observational data but not the data themselves, we alter the data interpretation process but not the data production process. I stress the contrast between production and interpretation in part because there is a strong tendency in philosophical discussion to suppose that improvements in the reliability of data-to-phenomena reasoning are achieved exclusively by changes in data interpretation and in particular the construction of better background theories. By contrast, in real science the most effective improvements in reliability very often are achieved by altering the data

production process—by building a better telescope or by dropping the fancy econometrics and doing an experiment instead.

3. Counterfactual Dependence and Reliability. This paper will explore some aspects of evidential reasoning from data to phenomena. The basic idea that I will be defending is that in many typical cases the relationship that must be present if data are to provide evidence for some phenomenon-claim is a certain sort of systematic pattern of counterfactual dependence or sensitivity of both the data and the conclusions investigators reach on the phenomena themselves. Suppose an investigator is considering a set of competing claims $P_1 \ldots P_n$ about some phenomenon of interest where these claims are mutually exclusive and exhaustive in the sense that they represent all of the various possibilities which are worth taking seriously. The data production process will be capable of producing a range of different data outcomes $D_1 \ldots D_m$. The interpretive component of the overall detection or measurement procedure will involve reaching conclusions about which of $P_1 \ldots P_n$ is correct on the basis of which of $D_1 \ldots D_m$ is produced. The interpretive component will thus specify a set of conditionals of the form: (1) If D_j is produced, conclude that P_i is true. Then the ideal at which one aims is: (2) the overall detection or measurement procedure should be such that each of the conditionals of form (1) recommends that one accept P_i when and only when P_i is correct. Somewhat more succinctly: the detection or measurement procedure should be such that different sorts of data $D_1 \ldots D_m$ are produced in such a way that investigators can use such data to reliably track exactly which of the competing claims $P_1 \ldots P_n$ is true.

Satisfaction of this ideal requires the right sort of counterfactual sensitivity in both the production and interpretive phases of the detection or measurement procedure; both must work together to insure that claims of form (2) are correct. In particular, one wants it to be the case both that different data $D_1 \ldots D_m$ are produced depending on which of the competing claims $P_1 \ldots P_n$ are true (the position of the pointer on the ammeter is dependent on or varies counterfactually with different values of the current) and that the conclusions that are recommended on the basis of the data track or are counterfactually sensitive to which of these claims is correct (the procedures for reading or interpreting the position of the pointer are such that we can infer the correct value of the current from the pointer position).

Use of the ammeter is naturally viewed as a case of *measurement*: one is interested in discriminating among a number of different values that the current may possess. We can think of cases of *detection* as a special case of this pattern in which there are just two possibilities—either a phenomenon is present in the experimental context (P_1) or it is not (P_2)—and two

corresponding data outcomes or sets of such outcomes D_1 and D_2. Associated with these will be conditionals of form (1) telling us whether to infer P_1 or P_2 depending on whether D_1 or D_2 is produced. What one wants of course is that these conditionals should tell us to infer P_1 (P_2) when and only when P_1 (P_2) is correct. As we shall see in more detail below, in the case of both measurement and detection, the extent to which this general pattern is satisfied will be a matter of degree—when or to the extent it is satisfied, I will say that the procedure and the associated evidential connection between data and phenomena are *reliable*.

The importance of the general sort of pattern of counterfactual dependence just described for evidential support is emphasized by a number of writers and methodological traditions. Closely related notions are employed by David Lewis (1986) to distinguish genuine seeing from (veridical or nonveridical) hallucination and by Robert Nozick (1981) in his tracking account of knowledge. As I note below, one can think of information about the error characteristics of repeatable methods, the role of which is emphasized in the reliabilist tradition in epistemology (Goldman 1986, Dretske 1981), in the Neyman-Pearson tradition in statistics, and in Deborah Mayo's important recent book, *Error and the Growth of Experimental Knowledge* (1996), as evidentially relevant because or to the extent that it furnishes grounds for belief in conditionals of form (2).

The demand that the relationship between data and competing phenomena claims exhibit the sort of pattern of counterfactual dependence described above is stronger than the demand that if data D_1 is to serve as evidence for some phenomenon claim P_1, then P_1 (or P_1 in conjunction with other plausible background assumptions) must be sufficient for D_1 or that the phenomenon described by P_1 must in fact cause D_1. Consider a case in which the presence of neutral currents is in fact sufficient for (in conjunction with true background assumptions) or has in fact caused certain bubble chamber photographs D_1. If some other competing phenomena claim P_2—e. g., that background neutrons are present—is also sufficient for D_1 or if background neutrons would also cause D_1 and if the experimental arrangement is such that the possibility that P_2 holds must be taken seriously, then the overall detection procedure is very likely such that one would conclude P_1 even if P_2 is correct. In this case the pattern of counterfactual dependence described above will not be satisfied and D_1 will not be good evidence for P_1.

The pattern of counterfactual dependence just described is an ideal which, for a variety of reasons is rarely fully satisfied in practice. For one thing, as we noted above, in typical cases data are the results of many causal factors and at most some of these will have to do with the phenomenon of interest. The operation of these additional causal factors will usually have the consequence that even in a well-designed experiment the

association between phenomena, data outcomes, and conclusions will be probabilistic rather than deterministic. A particular data outcome D_1 will sometimes occur in the presence of a range of different phenomena claims $P_1 \ldots P_n$, since one constellation of additional factors in conjunction with P_1 may produce D_1, and another constellation of additional factors in conjunction with P_2 may also produce D_1 and so on. For similar reasons the holding of a particular phenomenon claim P_1 will not be associated with a unique data outcome D_1 but rather with a range of such outcomes $D_1 \ldots D_i$, depending again on what additional factors are present. For example, in an experiment to measure the true melting point of some chemical compound (the phenomenon of interest), repeated thermometer readings (the data) will exhibit a scatter around the true melting point even if the thermometer is functioning properly, the measurements are made under appropriate conditions and so on. However, given certain assumptions about the distribution of these additional factors influencing the measurement result, some function of the individual measurement results such as their mean will have desirable features as an estimator of the true melting point. In cases of this sort, it is natural to think in terms of a probabilistic version of reliability: the detection procedure or measurement procedure will recommend accepting P_i on the basis of some observed set of data outcomes and one wants the procedure to be such that (3) the probability that it recommends inferring or accepting P_i given that P_i is true is high for each P_i, where the probabilities in question may be given a frequency or propensity interpretation. (The probabilities in question thus pertain to the error characteristics of the detection procedure that makes the recommendations rather than directly to the recommendations themselves). Both hypothesis testing and parameter estimation within a classical statistical framework as well as diagnostic medical tests when the associated error probabilities are known furnish familiar examples of this general idea.

In other cases in which the reliability of a detection or measurement technique is at issue, information about specific probability values associated with conditionals of form (3) may be lacking. Nonetheless, the general notion of reliability relevant to such techniques is recognizably similar to the notions described above: empirical investigation allows investigators to learn about the kinds of mistakes that do or do not occur with the repeated use of a measurement procedure in various applications and this is used to ground judgments of reliability. Consider, for example, the empirical investigations into the reliability of the potassium-argon dating method (Glen 1982). In the course of their investigations researchers learned empirically that the potassium-argon method was reliable with certain kinds of rocks and not others, that it was unreliable unless atmospheric argon was removed from the rock samples, and so on. It is perhaps

not too misleading to describe this as a matter of learning about the *qualitative error characteristics* of the potassium-argon method. That is, one might describe the reliability of a detection or measurement procedure like potassium-argon dating by means of a set of qualitative conditionals analogous to (3) which specify whether the procedure makes relatively few or many mistaken recommendations in various kinds of applications, even if one cannot attach precise numbers to the probabilities of these mistakes. A similar point holds for a number of other examples described below.

Another reason why the pattern of counterfactual dependence described above may not be achievable in practice is that it may not be possible to insure that conditionals of form (2) or (3) hold for all P_i. Consider an experimental procedure in which it is concluded on the basis of whether data D_1 or D_2 is produced that some phenomenon is either present (P_1) or absent (P_2). One may be able to design an experiment which has a high probability of concluding that P_2 holds when it does, but the price of doing this may be that the experiment has a low probability of concluding that P_1 holds when it does (the experiment is unlikely to falsely indicate that the phenomenon is present if it is not but has a high probability of failing to detect the phenomenon when it is present). Conversely, one may design an experiment which has a high probability of concluding that P_1 holds when it does (the experiment is unlikely to miss the phenomenon if it is present) but the price of this may be many false positives. In his well-known experiments to detect gravitational radiation, Joseph Weber preferred a nonlinear algorithm for the analysis of his data on the grounds that this was more likely than alternative algorithms to detect the sorts of pulses that he thought were characteristic of gravitational radiation, but as critics were able to show the algorithm was also more likely indicate the presence of gravitational radiation even though it was known on independent grounds to be absent (Franklin 1997, 49). A similar sort of tradeoff is present in many other experimental contexts and in many diagnostic tests.

Obviously, an acceptable detection procedure must have some positive probability of detecting the effect one is looking for, if that effect were to be present. For this reason Weber's critics, who had failed to detect gravitational radiation using an alternative experimental design and algorithm, went to considerable lengths to try to show their experiment was capable of detecting the sorts of pulses that are taken to be characteristic of gravitational radiation. Nonetheless, it is usually not necessary that a detection technique detect all instances of the phenomena of interest that occur within an experimental or measurement context— i.e., that data D_1 diagnostic of P_1 always occur whenever P_1 holds. Instead one needs to detect only enough apparent instances of P_1 to convince oneself and others that P_1 is genuine or to support reliable statistical inferences concerning P_1.

This is reflected in the trade off that is typically made between the two sorts of mistakes described above. As a general rule, researchers seem to attach more importance to avoidance of false claims that a phenomenon is present or has certain features than they do to avoiding failures to detect the phenomenon when present. Presumably this reflects a judgment of some sort about the relative epistemic or scientific costs of these sorts of mistakes—judgments that at least enter into experimental design even if they do not enter directly into assessments of evidential support. The role of such judgments in scientific investigation is very much under-explored in traditional philosophical accounts of confirmation.

4. Empirical Assessment of Reliability. How can investigators tell when the patterns of counterfactual dependence described above hold? In some cases one must rely on explicit derivation or calculation. To answer the question of whether the starlight deflection E observed by Eddington constitutes evidence for General Relativity we need to know, among other things, whether (4) E would hold if various alternatives to General Relativity were to hold instead and this requires in turn that we be able to formulate those alternatives and to calculate what they imply about starlight deflection. By contrast, in many cases of data-to-phenomena reasoning it is not necessary or even practically possible to determine whether the right sort of pattern of counterfactual dependence (conditionals of form (2) or (3)) holds by exclusive reliance on calculation and derivation. Instead investigators must rely on what I will call empirical investigation. This is more readily illustrated than defined but the general idea is that in assessing reliability in data-to-phenomena reasoning, scientists focus on a large number of highly specific local empirical facts about the causal characteristics of the detection or measurement process and investigate these by means of strategies (like calibration; see below) that need not involve explicit derivation. For example, scientific arguments about whether a particular experimental arrangement provides evidence for the existence of neutral currents will turn on whether various background factors that can mimic the effects of neutral currents have been adequately controlled for, whether those who sort through the bubble chamber photographs produced by the experiment can reliably distinguish those photographs that indicate neutral currents from those that are produced by other sorts of particle interactions, and so on. The empirical investigation, described above, of the circumstances in which potassium-argon dating is reliable, furnishes another example.

In describing such investigations as empirical, I do not mean to deny that they can be modeled or represented formally by philosophers or methodologists. My point is rather that the scientists who are engaged in such assessments of reliability do not themselves rely exclusively on logical or

formal structural or subject matter-independent relationships of the sort emphasized in traditional accounts of confirmation in assessing evidential support. For example, establishing that the potassium-argon dating technique is reliable is not a matter of deriving the characteristics of the data it produces from theory and initial conditions, as HD accounts would suggest. As we will see, scientists' ability to investigate the reliability of data-to-phenomena reasoning empirically means that distinctive epistemological strategies are available in connection with such reasoning that may have no obvious counterpart in cases in we are interested in the support phenomena claims provide for general theory, where heavy reliance on explicit derivation and logical relationships seems unavoidable. I think that it is this fact—that one can investigate the reliability of measurement and detection processes empirically—that constitutes the core of truth behind the claims of writers like Nozick (1981) and Achinstein (1991) that in many cases the evidential relationship is an empirical or factual rather than an a priori relationship. I interpret this to mean that formal models of evidential relationships should be understood as attempts to capture relationships that are grounded in empirical facts about the reliability characteristics of various detection processes.

As I have implied, the idea that the status of conditionals like (2) and (3) is an empirical issue is perhaps most plausible when one has a well-defined notion of what it is to repeatedly use the same detection or measurement procedure to generate a body of data and to reach conclusions regarding phenomena. In many such cases one can empirically investigate the error characteristics of the procedure under repetition and use this information to determine whether the appropriate sort of counterfactual dependence is present and to assess claims of sort (2) and (3). One important example of this is the strategy that Allan Franklin (1997) calls calibration. Here one assesses the error characteristics of a method by investigating its ability to detect or measure known phenomena and then assumes that the method will have similar error characteristics when used to investigate some novel phenomenon. Thus in an experiment to measure the infrared spectrum of an organic substance, the investigators knew in advance that the substance was prepared in a solvent and what the chemical composition of the solvent was. The ability of their apparatus to correctly measure the known spectrum of the solvent provided support for the claim that various possible sources of error had been controlled and that the apparatus also correctly measured the unknown spectrum, i.e., that in the case of the unknown substance a similar pattern of dependence of the data and the recommended conclusions on its actual spectrum were also satisfied. In this case, the reliability of the measurements for the unknown substance was not established by explicit calculation or derivation from some general theory of the measurement apparatus. Instead the ar-

gument for reliability rested on empirical facts about the behavior of the measurement apparatus in connection with the known sample and the empirical assumption that the error characteristics or discriminative abilities of the measurement procedure employed in the experiment were similar for both the known and unknown substances. The spurious positive signals generated by Weber's detection procedures (see above) are another case in which calibration is used empirically to assess reliability.

Although I lack the space for detailed discussion, many other strategies for establishing experimental reliability described by Franklin (1986), Hacking (1983), and Woodward (1989) can be viewed similarly as empirical arguments for assessing whether conditionals of form (2) and (3) are satisfied. Consider Hacking's well-known suggestion that one can argue for the reliability of the light microscope by manipulating the specimen in known ways and observing corresponding changes in the image. It is natural to think of this as an empirical investigation that confirms that a certain pattern of counterfactual dependence is present between features of the specimen and features of the image. One shows that this pattern of dependence is present by showing that when the specimen is changed in various ways the image changes in corresponding ways and this in turn entitles one to take the image as evidence for the state of the specimen. One can reason in this way without doing any explicit calculation or explicit statistical analysis at all.[1]

5. Logical Models of Confirmation. As I noted above, a number of writers (Achinstein 1991, Mayo 1996, Nozick 1981) have contrasted the idea that evidential support is at least sometimes an empirical matter with traditional philosophical models which see the evidential relationship as a

1. As argued in the text, when a determinate, repeatable experimental or measurement set-up is used to generate data relevant to discriminating among competing phenomena claims, talk of the error characteristics of the procedure under repetition will often make perfectly good sense. An interesting question is how far we can extend this paradigm. Once we distinguish between reasoning from data to phenomena and reasoning from phenomena to more general theoretical claims, we see that it is one thing to talk of the error characteristics of Eddington's procedures for inferring from his photographs to a value for the starlight deflection and another matter to talk about the error characteristics of some procedure involving the use of starlight deflection to test General Relativity. Does it makes sense to think in terms of a repeatable evidence generation or testing procedure with determinate empirically accessible error characteristics in the latter case as well as the former? As I understand the argument of Deborah Mayo's recent book (1996), her answer is "yes"; she wants to appeal to ideas about error characteristics to give an account of testing and evidence in science which is applicable quite generally and not just in the context of data-to-phenomena reasoning. While I share Mayo's emphasis on the importance of error characteristics in the context of data-to-phenomena reasoning, I confess to some uncertainty about how she proposes to extend these ideas to other contexts.

purely formal, logical, or a priori matter. Again, it is easier to give examples than to provide a precise characterization of such "logical" models, but the general idea is that when evidence supports a hypothesis this will be in virtue of the fact that the former stands to the latter in some inductive relationship that can be given a purely syntactic or structural or at least subject matter-independent characterization. Broadly speaking, the relation between evidence and hypothesis is regarded as being like the relationship between the premises and conclusion of a deductively valid argument, except of course that, unlike the latter, the former is supposed to be ampliative. In many of the most familiar models of this sort, the evidential relationship is characterized purely in terms of the resources of first-order deductive logic but in other cases the characterization may involve other sorts of formal or mathematical resources such as probability theory. Examples include various versions of hypothetico-deductivism and Hempelian positive instance theories. Carnapian inductive logic as well as some Bayesian treatments are very much in the spirit of logical models, but other Bayesian accounts are not[2].

A characteristic features of most of logical models is a commitment to the idea that the processes by which data is produced or generated and in particular facts about the way in which such processes determine not just which data is produced but which data could have been produced make no difference to the evidential significance of data. What is supposed to guarantee the goodness of the inference from evidence to hypothesis is simply the instantiation of the right sort of logical relationship between the evidence and hypothesis and not empirical facts about how the evidence was produced. (Analogously, the soundness of a deductive argument does not depend on how the premises come to be true.) Thus, for example, on a Hempelian positive instance conception of confirmation, evidence consisting of black ravens supports the hypothesis (H) that all ravens are black in virtue of its "logical" relationship to H, regardless of whether this evidence is produced by a machine that samples only black objects at random and examines them for whether they are ravens or instead randomly samples ravens and determines whether they are black. Colin Howson and Peter Urbach (1989, 171) rely on a similar idea about the irrelevance of the data production process when in the context of a discussion of stopping rules, they repeat the familiar criticism that "[a] significance test depends not only on the outcome that a trial produced but also on outcomes that it could have produced but did not." They take it as so obvious as not to require further argument that the

2. Whether a Bayesian account is logical in spirit depends very much on whether the account recognizes the importance of modeling the error characteristics of the data generation process. See fn. 4.

latter sort of information must be irrelevant to the evidential significance of the outcome that did occur.

In contrast to the position advocated by Howson and Urbach, it is an immediate consequence of the counterfactual characterization given above that we cannot tell whether D_1 is evidence for P_1 just by considering the relationship between D_1 and P_1 and whether D_1 occurs. Whether D_1 is evidence for P_1 also depends on whether alternative data to D_1 would be produced if P_2 were to hold instead, even if in fact D_1 are all the data that are ever produced. More generally, on the counterfactual characterization of evidence, the characteristics of the data production process are of crucial evidential significance because they influence which data outcomes would have been produced had different phenomena claims obtained and hence which conclusions would have been accepted regarding those claims. For example, whether certain bubble chamber photographs are evidence for neutral currents depends on whether the experimental set-up that produced them was such that even if neutral currents were absent exactly the same photographs that would have been produced (perhaps because of the presence of background neutrons) as if neutral currents were present and on whether the conclusion that neutral currents are present would have been drawn in both cases. If so, the photographs are not evidence for the presence of neutral currents. It thus follows that given two qualitatively identical or internally indistinguishable sets of data, one set may be evidence for some hypothesis and the other may not, depending on how they have been produced. Photographs that are internally indistinguishable from the original CERN photographs that were evidence for neutral currents may not themselves be evidence for neutral currents if, unlike the CERN photographs they were produced by a process that failed to control adequately for the presence of high-energy neutrons. At least in practice, logical models tend to neglect the possibility that internally indistinguishable bodies of data may differ in evidential significance in this way.[3]

I take it to be uncontroversial that, whatever their utility in illuminating how phenomena claims provide evidence for general theory, logical models of confirmation have proven to be rather unhelpful in understanding how data provide evidence for phenomena or how evidential reasoning works in experimental contexts. Why is this? Even a cursory acquaintance with the many detailed historical studies of experimentation that have emerged over the past decade or with the literature on experimental design

3. As Elliott Sober has pointed out to me, the observation that the evidential significance of data depends on the characteristics of the process that produces it may be regarded as an application of the more general point, made very persuasively in Good1967, that the evidential significance of data for a hypothesis depends not just on the data and hypothesis but upon which additional background assumptions are true in the situation under investigation.

and the design of observational studies shows that in virtually all areas of science information about the reliability of the processes by which data are produced or generated is taken to be crucial to its evidential significance. In my view, it is the neglect of such considerations in traditional logical models (and more generally a tendency to neglect the role of empirical considerations in assessing reliability, along with an exaggerated conception of the role played by explicit derivation in the assessment of evidential support) that accounts for the limited usefulness of such models in understanding how data provide evidence for phenomena.[4]

A natural response of defenders of traditional models is that it must always be possible in principle to make explicit the various background assumptions and theories that tell us whether an experimental apparatus is working properly, whether various sources of error have been controlled for, and so on, and that once we do so, the evidential relationship between data and phenomena can be characterized in formal a priori terms. The short answer to this suggestion is that whatever the force of "in principle" in the above claim, it is clear that scientists often do not actually provide, and are not in a position to provide, nontrivial derivations from independently-known premises that would establish that appropriate evidential connections hold between data and phenomena. As studies of past and contemporary science show, the sources of error that may infect a new experiment are so various, so difficult to detect and so local and idiosyncratic to the details of the experimental arrangement that often the only way to determine whether the experiment is working reliably (i.e., to determine whether it satisfies conditions of form (1) and (2)) are empirical investigations of the sorts described above. If we want to describe and assess the strategies of evidential reasoning scientists actually employ, con-

4. The dispute between Bayesians and advocates of classical (Neyman-Pearson) methods in statistics is often viewed, by both groups, as a dispute about the relevance of the error characteristics of repeatable processes to the assessment of evidence. In insisting on the relevance of the error characteristics of the data generation process, it may seem that I am siding with the latter tradition against the former. While I lack the space for detailed discussion, I am inclined to think that matters are more complex than this simple opposition suggests. What seems to follow from the argument given above is simply that when we have information about the error characteristics of well-defined repeatable processes involved in data generation and when such information is relevant to the assessment of conditionals of form (2) and (3), then adequate Bayesian treatments of the role of data as evidence will need to model or represent features of the data production process. In fact some Bayesians have explicitly advocated this (see, e.g., Pearl 1988, 58ff.) . The basic move is to incorporate information about the data generation process into the likelihoods or priors. Although some critics hold that there are reasons in principle why Bayesians cannot successfully capture the evidential significance of the data generation process, it is not part of my argument that this contention is correct.

siderations about what could be derived if investigators had unlimited computational powers and information that they do not possess are of dubious relevance.

6. Reliability, Circularity, and Derivability. The idea that one can often establish empirically that (4) a detection process is reliable without (5) deriving its reliability from some general theory of how that process works and/or why it is reliable is supported by a number of episodes in the history of science. For example, in a classic investigation into the differences in observations of stellar transit times reported by different astronomical observers, F. W. Bessel found systematic and relatively stable differences: "each observer has his own personal error" (Boring 1950, Gregory 1981). This made it possible to devise a "personal equation" relating the reported times for each observer and allowing different observers to be calibrated with respect to one another and to correct each other. Each observer was thus treated as an instrument with stable, long-run error characteristics, the reliability of which could be ascertained empirically. Bessel carried out this empirical investigation into the reliability of the human visual system in connection with a specific perceptual detection task even though he lacked (and to a large extent we still lack) a detailed explanatory theory of how the visual system works which would allow us to construct non trivial derivations of when it is reliable. Similarly, Galileo advanced a number of empirical arguments showing that his telescope was a reliable instrument in various astronomical applications even though he lacked a correct optical theory that could be used to explain how that instrument worked or why or when it was reliable. Donald Glaser invented the bubble chamber and showed it could be used reliably to detect subatomic particles, even though he and other investigators who used it were guided for several years by an incorrect theory of the mechanisms by which it worked (cf. Galison 1985)

Distinguishing between (4) and (5) also allows us to understand the acceptability of detection procedures and patterns of reasoning that otherwise might look objectionably circular. Consider an experiment to test an optical theory which involves a detection process in which human vision plays an important role—for example, the experiments Fresnel performed to test his version of the wave theory of light, in which he used his visual system to ascertain the position of various interference fringes. To explain in detail the contribution of Fresnel's visual system to reliable detection of the position of the fringes we must appeal to (among other things) a correct optical theory of the nature and transmission of light, for this is crucial to the detection process. However, which optical theory is correct is the very issue to which Fresnel's experiment is directed. If one holds that to show that a detection process is reliable one must possess a

theory which explains the workings of that process or which permits one to derive the conclusion that the process is reliable from a correct theory of its operation, it will be hard to avoid the conclusion that Fresnel's procedure is objectionably circular, that to establish the reliability of his detection process one must already know whether his optical theory is true. For example, this conclusion seems to follow from philosophical views (e.g., Kosso 1989) that demand that if an observation O that is to provide evidence for a theory T_1, any theory T_2 that "loads" or is presupposed by O must be independent of T_1, assuming that loads means something like "explains," as it clearly does for many adherents of this position. Distinguishing between (4) and (5) vindicates the commonsense judgment that there need be nothing objectionably circular about Fresnel's investigation. Empirical strategies like those described above can be used by Fresnel to establish the reliability of his visual system in detecting the positions of the fringes even in absence of a theory of how the visual system works. As long as Fresnel has good reason to think that his visual system can reliably detect the positions of the interference fringes, the fact that the theory under test is also the theory that (in part) explains the operation of that system or when and why it is reliable is irrelevant.

7. Data Selection. The distinction between (4) and (5) and the idea that it is typically phenomena rather than data that are potential objects of systematic explanation by general theory also has implications for how we should think about data selection. Decisions to ignore or discard data play a central role in virtually all data-to-phenomena reasoning and yet are neglected in many standard theories of confirmation. While the account developed above will not yield mechanical rules for when it is justifiable to discard data, it does at least make it intelligible what is involved in such decisions and why they are sometimes justifiable. It is legitimate to discard data when one has reason to believe that the process that has generated the data is unreliable, that counterfactuals of forms (2) and (3) fail to hold. Because there is no general requirement in science to explain or derive facts about data and because data are of scientific interest only insofar as they furnish reliable evidence about phenomena, unreliable data may be legitimately ignored—it need not be treated as a constraint on theorizing and there is no general requirement to track down all of the various causal factors (sources of error) that may have been at work in producing it. It is for just this reason that scientists usually do not publish (and do not regard themselves as under any methodological obligation to publish) all their data but (at best) only that portion of it which is reliable evidence.

We may contrast this way of thinking about data selection with the very different attitude that is naturally suggested by conceptions that take the characteristics of data generation processes to be irrelevant to the ev-

idential significance of data and which regard data rather than phenomena to be the primary objects of scientific explanation. Because decisions to discard data are based on researchers' judgments about the characteristics of the data production process, logical models of evidential support will find it difficult to provide a plausible treatment of such decisions and will tend to regard them with suspicion, especially when the causal factors involved in the production of the discarded data remain unknown and we have no independent evidence that various relevant background assumptions are or are not satisfied in connection with this data (see below). Similarly, such conceptions will find it hard to explain why scientists should not recognize a general requirement to publish and explain all data, whether or not it is the result of a reliable data production process, so that the causal factors underlying the production of unreliable data must always be tracked down.

While I lack the space for detailed discussion, I believe that one can see the influence of this second attitude toward data selection in a number of disputes about misconduct in science—for example, among those who fault Millikan in his 1913 account of his oil drop experiment for failing to publish all of his data and for discarding data even though he was unable to establish on independent grounds which sources of error were present. By contrast Millikan's conduct is more understandable on the picture of data-to-phenomena reasoning defended in this essay. Data are the product of a large number of disparate causal factors, many of which may be difficult to identify. For this reason, it is extremely common for experimental or measurement procedures to fail to be reliable for unknown reasons and in such cases scientists often do not concern themselves with tracking down the exact source of the error; this is regarded as both difficult to do and as having little scientific payoff, since explaining data per se is not a goal of inquiry and since data are of scientific interest and a constraint on theorizing only insofar as they provide evidence for phenomena. This is not to say, of course, that an investigator's freedom to discard data should be utterly unconstrained, but rather that there will be an important and probably ineliminable role for judgment (including judgments about the relative importance of various sorts of mistakes) in such matters and that as long as discarding data turns out in fact to enhance or at least not undermine the reliability of the experiment, as in fact was the case with Millikan, it shouldn't be subject to methodological criticism, even if the researcher is unable to provide a compelling independent justification for such decisions.

REFERENCES

Achinstein, Peter (1991), *Particles and Waves: Historical Essays in the Philosophy of Science.* Oxford: Oxford University Press.

Bogen, James and James Woodward (1988), "Saving the Phenomena", *The Philosophical Review* 97: 303–352.
———. (1992), "Observations, Theories and the Evolution of the Human Spirit", *Philosophy of Science* 59: 590–611.
———. (forthcoming). "Evading the IRS", in N. Cartwright and M. Jones (eds.), *Correcting the Model: Idealization and Abstraction in Science.* Amsterdam: Rodolpi.
Boring, E. (1950), *A History of Experimental Psychology,* 2nd ed. New York: Appleton-Century-Crofts.
Dretske, F. (1981), *Knowledge and the Flow of Information.* Cambridge, MA: MIT Press.
Franklin, Allen (1986), *The Neglect of Experiment*. Cambridge: Cambridge University Press.
———. (1997) " Calibration", *Perspectives on Science* 5: 31–80.
Galison, P. (1985), "Bubble Chambers and the Experimental Workplace", in P. Achinstein and O. Hannaway (eds.), *Observation, Experiment and Hypothesis in Modern Physical Science*. Cambridge, MA: MIT Press, 309–373.
Glen, William (1982), *The Road to Jamarillo*. Stanford: Stanford University Press.
Goldman, Alvin (1986), *Epistemology and Cognition.* Cambridge, MA: Harvard University Press.
Good, Irving (1967), " The White Shoe is a Red Herring", British Journal for the Philosophy of Science 17: 322.
Gregory, Richard (1981), *Mind in Science: A History of Explanations in Psychology and Physics*. Cambridge: Cambridge University Press.
Hacking, Ian (1983), *Representing and Intervening.* Cambridge: Cambridge University Press.
Howson, Colin and Peter Urbach (1989), *Scientific Reasoning: The Bayesian Approach.* La Salle, IL: Open Court.
Kosso, P. (1989),"Science and Objectivity", *The Journal of Philosophy* 86: 245–257.
Lewis, D. (1986), "Veridical Hallucination and Prosthetic Vision", reprinted in D. Lewis, *Philosophical Papers,* vol. 2. New York: Oxford University Press, 000–000.
Mayo, Deborah (1996), *Error and the Growth of Experimental Knowledge.* Chicago: University of Chicago Press.
Millikan, R. (1913), "On the Elementary Charge and the Avogadro Constant", *Physical Review* 2: 109–143.
Nozick, Robert (1981), *Philosophical Explanations. Cambridge*, MA: Harvard University Press.
Pearl, Judea (1988), *Probabalistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan and Kauffman.
Woodward, James (1989), "Data and Phenomena", *Synthese* 79: 393–472.