Poznań Studies in the Philosophy of the Sciences and the Humanities
2003, vol. XX, pp. 223-256

James Bogen and James Woodward

EVADING THE IRS

I

"IRS" is our term for a view about theory testing originated by members and associates of the Vienna Circle. Its leading idea is that the epistemic bearing of observational evidence on a scientific theory is best understood in terms of Inferential Relations between Sentences which represent the evidence and sentences which represent hypotheses belonging to the theory. The best known versions of IRS (and the ones we concentrate on in this paper) are Hypothetico-Deductive and positive instance (including bootstrapping) confirmation theories. It goes without saying that such accounts, along with the problems they generate, have exerted a dominant influence on philosophers who study the epistemology of science.

We maintain that the epistemic import of observational evidence is to be understood in terms of empirical facts about particular causal connections and about the error characteristics of detection processes. These connections and characteristics are neither constituted by nor greatly illuminated by considering the formal relations between sentential structures which IRS models focus on. We argue that by taking them seriously, you too can evade the IRS.

We have argued elsewhere[1] that theory testing in the natural and social sciences is typically a two-stage process and that the use of observational evidence belongs primarily to the first stage. In this stage data are produced and interpreted in order to draw conclusions about what we call phenomena. This is usually a matter of considering a number of competing claims about the phenomenon under investigation and using the data to decide which of those claims is most likely to be correct. In the second stage, theoretical claims are confronted with conclusions about phenomena reached in the first stage. Some examples of data are records of temperature readings used to determine the melting point of a substance, scores on psychological tests used to investigate memory processing, bubble chamber and spark detector photographs used to

[1] Bogen and Woodward (1988); Woodward (1989); Bogen and Woodward (1992).

detect particle interactions, drawings of prepared tissue viewed under microscopes used to determine the structure of neural systems, and the eclipse photographs Eddington, Curtis and others used to calculate the deflection of starlight by the sun. Some examples of phenomena are the melting points calculated from the temperature readings, the widespread and regularly occurring features of memory processing investigated through the use of psychological tests, the deflection of starlight, etc. Data are effects produced by elaborate causal processes that may involve the operation of the human perceptual and cognitive systems as well as measuring and recording devices and many other sorts of natural and manufactured non-human systems. We think the epistemic importance of the human perceptual system in data production depends upon its influence on the reliability of the procedures by which the data are produced and interpreted. In this respect there is no epistemically interesting difference between human perception and any other factor which influences reliability.[2]

The epistemic significance of data depends upon whether they possess features through which the phenomena of interest can be studied. It depends also upon whether they can be inspected and analyzed by investigators who wish to use them. The production of data meeting both of these conditions usually requires the manipulation of highly transitory and unusual combinations of causal factors which do not naturally operate together in any regular way. In many cases these causal structures are idiosyncratic to highly unusual situations many of which are highly contrived and peculiar to the laboratory. In contrast, phenomena are typically due to the uniform operation of a relatively small number of factors whose operation does not depend upon the rare and often highly artificial settings required for data production. As a result, many phenomena are capable of occurring in a variety of different natural and contrived settings.[3]

It is phenomena rather than data that scientists typically seek to explain and predict. We believe that in most cases, scientific theories are tested directly against phenomena rather than data. For example Einstein's theory of general relativity was tested against a value for the deflection of starlight, rather than the photographs from which the deflection was calculated. The electro-weak theory devised by Weinberg and Salam was tested against claims about a phenomenon (the occurrence of neutral currents) rather than against the bubble chamber and spark detector data on which those claims were based. The testing of Newton's theory of universal gravitation involved such phenomena claims as Kepler's and Galileo's laws rather than the data used to investigate these phenomena (e.g., descriptions of pendulum and inclined plane experiments, astronomical records of the movement of the moon, etc.). The second stage of theory testing is the confrontation of theory with phenomena.

---

[2] Bogen and Woodward (1992).

[3] Bogen and Woodward (1988), p. 319ff.

We will use the term "observation sentences"[4] in connection with the sentences (also called "protocol sentences," "evidence sentences," etc.) the IRS uses to represent empirical evidence. Although details vary and controversy abounds, the IRS literature tends to associate them with reports of what individual observers perceive. As will be seen in sections IV and V below, it is often very hard to see how to construct a sentence which both represents the photographs and other non-sentential evidence scientists often use as data and also captures what is epistemically significant about them. Nevertheless we assume the IRS notion of an observation sentence was meant to play something like the same role as our notion of data; both notions are meant to explain the role of empirical evidence in theory testing. Accordingly, we shall speak of observation sentences as "corresponding" to data, but with the caveat (to be illustrated in section IV) that the details of this correspondence are often quite unclear.

By picturing theory testing as a one-step confrontation of theory with the evidence which "observation sentences" are meant to represent, the IRS ignores the two-tiered structure just sketched. And as the bulk of this paper will be devoted to suggesting, we think the IRS picture does not provide an adequate account of real world scientific reasoning from data to phenomena.

Our own view is that with regard to the investigation of phenomena, the evidential value of data is assessed in terms of general and local reliability. As we use these terms, *general reliability* depends upon the long-run error characteristics of repeatable processes for data production and interpretation.[5] We discuss it in section VII below. Generally reliable detection procedures may fail, and generally unreliable procedures may succeed in enabling an investigator to discriminate correctly in a particular case.[6] Furthermore, some procedures used in one or a very few cases are not, or cannot be repeated as would be needed to establish their long-term error characteristics. *Local*

---

[4] In this we depart from the IRS literature in which the term "observation sentence" is used for natural language observation reports as well as their counterparts in first order logical languages. We emphasize that we are using the term only for the latter.

[5] This is roughly the notion invoked by Alvin Goldman and other reliabilists. See, e.g., Goldman (1986), chs. 4, 5, 9-15 *passim*. However, unlike Goldman, we think, for reasons that will emerge in section IX, that the project of investigating the reliability characteristics of most human belief-forming methods and mechanisms is unlikely to be illuminating or fruitful. Rather we apply the notion of general reliability to highly specific measurement and detection procedures, or in connection with the use of instruments for particular purposes. Such procedures and uses of instruments often have determinate error characteristics that we know how to investigate empirically, while we suspect that this is not true of many of the methods or psychological processes that underlie belief formation.

[6] For example, consider a technique for staining tissue to be viewed under a light microscope which (like golgi staining) tends to produce a great many artifacts. The staining technique may nevertheless occasionally produce preparations which are free from artifacts, or whose artifacts can be easily distinguished from real cell structures. In such cases a generally unreliable microscopic technique can be locally reliable, and recognizably so.

reliability has to do with single case performances of procedures, pieces of equipment, etc. We discuss this in section VIII. We will argue that neither general nor local reliability can be assessed by considering the data all by itself without considering the processes by which it was produced and interpreted. These processes are the loci of the empirical facts upon which both the local and the general reliability – and hence, the evidential value – of data often depends. The last section of our paper argues that IRS neglects and lacks the resources needed to deal informatively with these epistemically crucial factors.[7]

II

As compared to the best studies by historians, sociologists, and anthropologists of science, the IRS literature contains little that can be easily recognized as belonging to actual scientific practice. While IRS analyses rely heavily on a logical formalism which is known by few and used by fewer practicing scientists, the mathematical formalisms natural scientists actually rely upon do little work in the IRS literature.[8] More importantly many data consist of photographs, drawings, tables of numbers, etc., which are not at all sentential in form, and scientific hypotheses are almost always set out in languages which are very different from first order logic. In contrast, the versions of IRS we consider try to account for the evidential relevance of data to theoretical claims in terms of a confirmation relation (see III below) characterized in terms of relations between sentences in a first order language. All of this is remarkable enough to raise questions about what motivates the IRS. The following motivational sketch is intended to indicate why this program might have seemed worth pursuing, and also to show that striking discrepancies between scientific practice and its IRS depiction derive non-accidentally from its basic goals and strategies.

Like many of its founders and proponents Hempel saw the IRS as an alternative to the idea that scientific theories are not or cannot be tested

---

[7] The relevance of causal factors in assessing evidential significance is also emphasized in Miller (1987). While we find much that is valuable and insightful in Miller's discussion, his account diverges from ours in important respects – in particular he tends to see inductive inference generally as a species of inference to the best explanation, while we do not. The evidential relevance of data-generating processes and the limitations of formal accounts of evidential support are also emphasized in Humphreys (1989), a discussion we have found very helpful.

[8] For an excellent and forceful characterization of the disparity between the literature of science and the literature of IRS-influenced philosophy of science, see Feyerabend (1985), pp. 83-85. Although we disagree with much of what Feyerabend says elsewhere we heartily endorse his idea in this passage that much of what occupies the IRS philosophers is an artifact of their own picture of science, and in particular, that much (Feyerabend would probably say "nearly all") research in the philosophy of science "consists in proposing ideas that fit the boundary conditions, i.e., the standards of the simple logic" chosen by the logical positivists to represent scientific reasoning (*ibid*., p. 85).

objectively – that "the decision as to whether a given hypothesis is acceptable in the light of a given body of evidence" rests on nothing more than a subjective "'sense of evidence,' or a feeling of plausibility in view of the relevant data." This, says Hempel, is analogous to the equally noxious idea that "the validity of a mathematical proof or of a logical argument has to be judged ultimately by reference to a feeling of soundness or convincingness." Hempel thinks both ideas rest on a confusion of rational, objective, logical factors which can actually determine whether the available evidence warrants the acceptance or rejection of a scientific hypothesis with subjective psychological factors which may influence scientific belief. To disentangle them we need purely formal criteria for confirmation of the kind deductive logic provides for the validity of deductive arguments. Such criteria would provide for "rational reconstruction[s] of the standards of scientific validation," free from the influence of feelings of conviction, senses of evidence, or other subjective factors which vary "from person to person, and with the same person in the course of time." And like the standards by which deductive validity is judged, "it seems reasonable to require that the criteria of empirical confirmation, besides being objective in character, should contain no reference to the specific subject matter of the hypothesis or of the evidence in question."[9]

The application of this approach to a real life example of scientific reasoning from evidence to a conclusion begins with the construction of a highly idealized representation of the reasoning under consideration. Reichenbach describes this as the construction of a "logical substitute" for the "real processes" by which the scientist thinks about the evidence.[10] As he describes it, this is analogous to replacing an informal deductive argument with a formal version which omits logically irrelevant features and exhibits logical structure which was not explicit in the original version. For Hempel, it is analogous to the construction of an idealized, simplified theoretical model of a real process.[11]

Once a rational reconstruction of a particular argument from evidence has been produced, the next step in an IRS treatment is the application of logical standards (Hempel's "objective criteria") to the reconstruction. This is analogous to applying logical rules to the formalized version of an argument to

---

[9] Hempel (1965), pp. 9-10. This is exactly what Glymour promises for his bootstrap theory. Its confirmation relations are to be "entirely structural; they have no connection to the content of the hypothesis tested, or to the meaning of the evidence sentences, or to the meaning of the theories with respect to which the tests are supposed to be carried out" (Glymour, 1980, pp. 374-5). The goal shared by Hempel and Glymour is closely related to Popper's goal of providing as formal as possible a demarcation between real and pseudo science. And it bears an interesting relation to Kuhn, Feyerabend, Hanson, Shapere, Quine, and many other critics of the original positivist program. Different as their views obviously are, all of these people subscribe to some version of the idea that the IRS is the only alternative to the idea that scientific belief is not objectively constrained by evidence.

[10] Reichenbach (1938), p. 5.

[11] Hempel (1965), p. 44.

explain whether (and under what interpretations) its conclusion is well supported by its premises. It is also analogous to explaining aspects of the behavior of a real system by appeal to the behavior of the items in a theoretical model.

The pursuit of these analogies made it natural if not inevitable for the IRS to leave out a great deal of what seems to us to be most characteristic of real world scientific testing. Thus Reichenbach insists it is no objection to his program that its "fictive constructions" do not "correspond at every point" to the actual thought processes of working scientists.[12] We respect Reichenbach's point: discrepancies between an idealization and a real system constitute serious objections only insofar as they defeat the purpose for which the idealization is used.[13] But we think the formally defined confirmation relations of the IRS fail to correspond to the evidential relevance of data to theory in ways which render the IRS picture uninformative in many cases, and seriously misleading in others.

III

The versions of IRS we will discuss are positive instance (including bootstrapping) accounts and Hypothetico-Deductive (HD) accounts of theory testing.[14] Their models are populated by sentences of a first order language. As noted, observational evidence is represented by observation sentences. Theoretical claims under test are represented by what we will call "hypothesis sentences." The resources of first order logic are used to characterize a relation of evidential relevance called "confirmation." Although observational evidence is said to "confirm" hypotheses or theories, the obtaining of the confirmation relation depends upon logical relations between what we are calling observation and hypotheses sentences. Simple versions of HD depict the confirmation of the theoretical claim corresponding to a hypothesis sentence, *h*, by evidence represented by an observation sentence, *o*, as depending on whether *(h & A)* deductively entails *o*. Here *A* is a first order representation of one or more "auxiliary hypotheses," "correspondence rules," or "background beliefs" which belong to the same theory as the claim represented by *h*. The simplest positive instance versions of IRS characterize confirmation in terms of logical relations which run in exactly the opposite direction. Where evidence represented by *o*

---

[12] Reichenbach (1938), p. 6.

[13] Thus Reichenbach requires the "construction...[to be] bound to actual thinking by the postulate of correspondence" (*ibid*.) and Hempel says the model should conform to actual behavior as far as it can without violating constraints imposed for the sake of attaining "simplicity, consistency, and comprehensiveness" (Hempel, 1965, p. 44).

[14] For positive instance accounts, see "Studies in the Logic of Confirmation" in Hempel (1965), pp. 3-46. For bootstrapping accounts, see Glymour (1980). For a simple HD account see Braithwaite (1953) and Popper (1959). For more complex HD accounts see Schlesinger (1976) and Merrill (1979).

confirms *h*, *o* (or, in the bootstrap version, the conjunction of *o* and *A*) entails an instance of *h*.[15]

Just as entailment can hold between false as well as true sentences, confirmation can relate worthless evidence to unacceptable hypotheses as well as good evidence to correct or well justified theoretical claims. Just as the mere fact that *p* entails *q* does not tell us whether we should believe *q*, the mere fact that *o* stands in the required inferential relation to *h* does not tell us whether there is good reason to accept the claim *h* represents. So what can IRS tell us about the acceptance and rejection of theoretical claims? Let a broken arrow (--->) represent the inferential relation used to characterize confirmation. A naive HD answer to our question would be that if o ---> *h*, evidence which makes *o* true provides epistemic support for the claim represented by *h* or the theory to which that claim belongs and evidence which makes ~*o* true provides epistemic support for the rejection of the claim represented by *h*. A simplified positive instance answer would be that the evidence represented by *o* provides epistemic support for the claim represented by *h* if *o* is true and *o* ---> *h*, while evidence counts against the claim if *o* is true and *o* ---> ~*h*.[16]

IV

We have emphasized that the IRS depicts confirmation as depending upon formal relations between sentences in a first order language, even though many data are photographs, drawings, etc., which are not sentences in any language let alone a first order one. This is enough to establish that the claim that confirmation captures what is essential to evidential relevance is not trivial. In fact that claim is problematic. Hempel's raven paradox illustrates one of its problems. Replacing the natural language predicate "is a raven" with *F*, and "is black" with *G*, let a hypothesis sentence (h_1), (x) $(Fx \supset Gx)$, represent the general claim (*C*) "All ravens are black."[17] Where *a* is a name, *Fa* & *Ga* is an instance of (h_1). But (h_1) is logically equivalent to (h_2), $(x)(\sim Gx \supset \sim Fx)$. Now ~*Fa* & ~*Ga* entails ~*Ga* ⊃ ~*Fa*, an instance of (h_2). Thus, ~*Fa* & ~*Ga* ---> (h_2). But as Hempel observes, ~*Fa* & ~*Ga* is true when the referent of *a* is a red

---

[15] Different versions of HD and positive instance theories add different conditions on confirmation to meet counterexamples which concern them. For example, o may be required to have a chance of being false, to be consistent with the theory whose claims are to be tested, to be such that its denial would count against the claim it would support, etc. The details of such conditions do not affect our arguments. Thus our discussion frequently assumes these additional conditions are met so that its being the case that *o* ---> *h* is sufficient for confirmation of the claims represented by *h* by the evidence represented by *o*.

[16] See the previous note. For examples of this view, see Braithwaite (1953) and Glymour (1980), ch. V.

[17] Examples featuring items which sound more theoretical than birds and colors are easily produced.

pencil.[18] Therefore, assuming that evidence which confirms a claim also confirms claims which are logically equivalent to it, why shouldn't the observation of a red pencil confirm (*C*)? If it does, this version of IRS allows evidence (e.g., red pencil observations) to confirm theoretical claims (like "All ravens are black") to which it is epistemically quite irrelevant. Since the premises of a deductively valid argument cannot fail to be relevant to its conclusion, this (along with such related puzzles as Goodman's grue riddle and Glymour's problem of irrelevant conjunction (Glymour, 1980, p. 31) points to a serious disanalogy between deductive validity and confirmation. While the deductive validity of an argument guarantees in every case that if its premises are true, then one has a compelling reason to believe its conclusion, the evidence represented by *o* can be epistemically irrelevant to the hypothesis represented by *h* even though *o* ---> *h*.

The most popular response to such difficulties is to tinker with IRS by adding or modifying the formal requirements for confirmation. But close variants of the above puzzles tend to reappear in more complicated IRS models.[19] We think this is symptomatic of the fact that evidential relevance depends upon features of the causal processes by which the evidence is produced and that the formal resources IRS has at its disposal are not very good at capturing or tracking these factors or the reasoning which depends upon them. This is why the tinkering doesn't work.

An equally serious problem emerges if we consider the following analogy: Just as we can't tell whether we must accept the conclusion of a deductively valid argument unless we can decide whether its premises are true, the fact that *o* ---> *h* doesn't give us any reason to believe a theoretical claim unless we can decide whether o is true. To see why this is a problem for the IRS consider the test Priestley and Lavoisier used to show that the gas produced by heating oxides of mercury, iron, and some other metals differ from atmospheric air.[20] Anachronistically described, it depends on the fact that the gas in question was oxygen and that oxygen combines with what Priestley called "nitrous air" (nitric oxide) to produce water-soluble nitrous oxide. To perform the test, one combines measured amounts of nitric oxide and the gas to be tested over water in an inverted graduated tube sealed at the top. As the nitrous oxide thus produced dissolves, the total volume of gas decreases, allowing the water to rise in the tube. At fixed volumes, the more uncompounded oxygen a gas contains, the greater will be the decrease in volume of gas. The decrease is measured by watching how far the water rises. In their first experiments with this test, Priestley and Lavoisier both reported that the addition of "nitrous air" to the unknown gas released from heated red oxide of mercury decreased the volume

---

[18] Hempel (1965), p. 15f. Cf. Glymour (1980), p. 15ff.

[19] For an illustration of this point in connection with Glymour's treatment of the problem of irrelevant conjunction, see Woodward (1983).

[20] This example is also discussed in Bogen and Woodward (1992).

of the latter by roughly the amount previously observed for atmospheric air. This datum could not be used to distinguish oxygen from atmospheric air. In later experiments Priestley obtained data which could be used to make the distinction. When he added three measures of "nitrous air" to two measures of the unknown gas, the volume of gas in the tube dropped to one measure. Lavoisier eventually "obtained roughly similar results."[21] The available equipment and techniques for measuring gases, for introducing them into the graduated tube, and for measuring volumes were such as to make it impossible for either investigator to obtain accurate measures of the true decreases in volume.[22] Therefore an IRS account which thinks of the data as putative measures of real decreases should treat observation sentences representing the data from the later as well as from the earlier experiments as false. But while unsound deductive arguments provide no epistemic support for their conclusions, the inaccurate data from Priestley's and Lavoisier's later experiments provide good reason to believe the phenomena claim for which they were used to argue.

Alternatively, suppose the data were meant to report how things looked to Priestley and Lavoisier instead of reporting the true magnitudes of the volume decreases. If Priestley and Lavoisier were good at writing down what they saw, observation sentences representing the worthless data from the first experiments should be counted as true along with observation sentences representing the epistemically valuable data from the later experiments. But while all deductively sound arguments support their conclusions, only data from the later experiments supported the claim that the gas released from red oxide or mercury differs from atmospheric air. Here the analogy between deductive soundness and confirmation by good evidence goes lame unless the IRS has a principled way to assign true observation sentences to the inaccurate but epistemically valuable data from the later experiments, and false observation sentences to the inaccurate but epistemically worthless data from the earlier experiments. If truth values must be allocated on the basis of something other than the accuracy of the data they represent, it is far from clear how the IRS is to allocate them.

To avoid the problem posed by Priestley's and Lavoisier's data the IRS must assign true observation sentences to epistemically good evidence and false ones to epistemically bad evidence. What determines whether evidence is good or bad? The following example illustrates our view that the relevance of evidence to theory and the epistemic value of the evidence depends in large part upon causal factors. If we are right about this, it is to be expected – as we will argue in sections VII and VIII – that decisions about the value of evidence depend in large part upon a sort of causal reasoning concerned with what we are calling reliability.

---

[21] See Conant (1957); Lavoisier (1965), pt. I, chs. 1-4.
[22] Priestley (1970), pp. 23-41.

V

Curtis and Campbell, Eddington and Cottingham (among others) produced astronomical data to test Einstein's theory of general relativity. One of the phenomena general relativity can be used to make predictions about is the deflection of starlight due to the gravitational influence of the sun. Eddington and the others tried to produce data which would enable investigators to decide between three competing claims about this phenomenon: (N) no deflection at all, (E) deflection of the magnitude predicted by general relativity, and (NS) deflection of a different magnitude predicted by Soldner from Newtonian physics augmented by assumptions needed to apply it to the motion of light.[23] (N) and (NS) would count against general relativity while (E) would count in favor of it. The data used to decide between these alternatives included photographs of stars taken in daytime during a solar eclipse, comparison photographs taken at night later in the year when the starlight which reached the photographic equipment would not pass as near to the sun, and check photographs of stars used to establish scale. To interpret the photographs, the investigators would have to establish their scale, i.e., the correspondence of radial distances between stars shown on an accurate star map to linear distances between star images on the photographs. They would have to measure differences between the positions of star images on the eclipse and the comparison photographs. They would have to calculate the deflection of starlight in seconds of arc from displacements of the star images together with the scale. At each step of the way they would have to correct for errors of different kinds from different sources.[24]

The evidential bearing of the photographic data on Einstein's theory is an instance of what IRS accounts of confirmation are supposed to explain. This evidential bearing depended upon two considerations: (1) the usefulness of the data in discriminating between phenomena claims (N), (NS), (E), and (2) the degree to which (E), the value predicted by general relativity, disagrees with predictions based on the competitor theories under consideration. (1) belongs to the first of the two stages of theory testing we mentioned in section I: the production and interpretation of data to answer a question about phenomena. (2) belongs to the second of these stages – the use of a phenomena claim to argue for or against part of a theory. With regard to the first of these considerations, evidential relevance depends upon the extent (if any) to which differences between the positions of star images on eclipse and comparison pictures are due to differences between paths of starlight due to the gravitational influence of the sun. Even if the IRS has the resources to analyze

---

[23] Soldner's is roughly the same as a value predicted from an earlier theory of Einstein's. SeePais (1982), p. 304.

[24] Earman and Glymour (1980), p. 59.

the prediction of (E) from Einstein's theory, the relevance of the data to (E) would be another matter.[25]

Assuming that the sun's gravitational field is causally relevant to differences between positions of eclipse and comparison images, the evidential value of the data depended upon a great many other factors as well. Some of these had to do with the instruments and techniques used to measure distances on the photograph. Some had to do with the resources available to the investigator for deciding whether and to what extent measured displacements of star images were due to the deflection of starlight rather than extraneous influences. As Fig. 1 indicates, one such factor was change in camera angle due to the motion of the earth. Another was parallax resulting from the distance between the geographical locations from which Eddington's eclipse and comparison pictures were taken.[26]

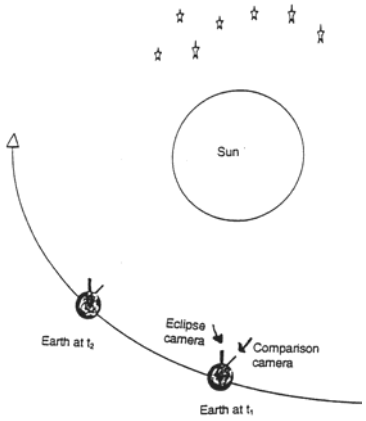


Fig. 1. As the earth moves from its position at one time, t_1, to its position at a later time, t_2, the positions of the eclipse and comparison cameras change relative to the stars.

---

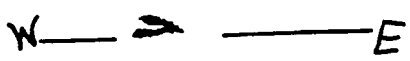
[25] In the discussion which follows, we ignore the fact that the deflection values calculated from the best photographic data differed not only from (N) and (NS), but also (albeit to a lesser extent) from (E). Assuming (as we do) that the data supported general relativity this might mean that although (E) is correct, its discrimination does not require it to be identical to the value calculated from the photographs. Alternatively, it might mean that (E) is false, but that just as inaccurate data can make it reasonable to believe a phenomenon-claim, some false phenomena claims provide epistemically good support for theoretical claims in whose testing they are employed. Deciding which if either of these alternatives is correct falls beyond the scope of this paper. But since epistemic support by inaccurate data and confirmation by false claims are major difficulties for IRS, the disparities between (E) and magnitudes calculated from the best data offer no aid and comfort to the IRS analysis. Important as they are in connection with other epistemological issues, these disparities will not affect the arguments of this paper.

[26] Eddington and Cottingham took eclipse photographs from Principe, but logistical complications made it necessary for them to have comparison pictures taken from Oxford. In addition to correcting for parallax, they had to establish scale for photographs taken from two very different locations with very different equipment (Earman and Glymour, 1980, pp. 73-4).

Apart from these influences, a number of factors including changes in temperature could produce mechanical effects in the photographic equipment sufficient to cause significant differences in scale.[27]

Additional complications arose from causes involved in the process of interpretation. One procedure for measuring distances between star images utilizes a low power microscope equipped with a cross hair. Having locked a photograph onto an illuminated frame, the investigator locates a star image (or part of one) against the cross hair and slowly turns a crank until the image whose distance from the first is to be measured appears against the cross hair. At each turn of the crank a gadget registers the distance traversed by the microscope in microns and fractions of microns. The distance is recorded, the photograph is removed from the frame, and the procedure is repeated with the next photograph.[28] If the photographs are not oriented in the same way on the frame, image displacements will be measured incorrectly.[29]

The following drawing of a star image from one of Curtis's photographs[30] illustrates effects (produced by the focus and by motions of the camera) which make this bit of data epistemically irrelevant to the testing of general relativity theory by rendering it useless in deciding between (E), (N), and (NS).



The epistemic defects of Curtis's star image are not due to the failure of inferential connections between an observation sentence and a hypothesis sentence. Nor are they due to the falsity of an observation sentence. By the same token, the epistemic value of the best photographs was not due to the truth of observation sentences or to the obtaining of inferential connections between them and hypothesis sentences. The evidential value of the starlight data depended upon non-logical, extra-linguistic relations between non-sentential features of photographs and causes which are not sentential structures.

At this point we need to say a little more about a difficulty we mentioned in section I. Observation sentences are supposed to represent evidence. But the IRS tends to associate evidence with sentences reporting observations, and even though some investigations use data of this sort, the data which supported (E) were not linguistic items of any sort, let alone sentences. They were photographs. This is not an unusual case. So many investigations depend upon non-sentential data that it would be fatal for the IRS to maintain that all scientific evidence consists of observation reports (let alone the expressions in first order

---

[27] Earman and Glymour (1980).

[28] We are indebted to Alma Zook of the Pomona College physics department for showing and explaining the use of such measuring equipment to us.

[29] Earman and Glymour (1980), p. 59.

[30] From a letter from Campbell to Curtis, reproduced in Earman and Glymour (1980), p. 67.

logic we are calling observation sentences). What then do observation sentences represent? The most charitable answer would be that they represent whatever data are actually used as evidence, even where the data are not observation reports. But this does not tell us which observation sentences to use to represent the photographs. Thus a serious difficulty is that for theory testing which involves non-sentential evidence, the IRS provides no guidance for the construction of the required observation sentences.

Lacking an account of what observation sentences the IRS would use to represent the photographs, it is hard to talk about what would decide their truth values. But we can say this much: whatever the observation sentences may be, their truth had better not depend upon how well the photographs depicted the true positions of the stars. The photographs did not purport to show (and were not used to calculate) their actual positions or the true magnitudes of distances between them. They could represent true positions of (or distances between) stars with equal accuracy only if there were no significant[31] discrepancies between the positions of star images on the eclipse photographs and star images on the comparison photographs. But had there been no such discrepancies the photographs would have argued against (E). Thus to require both the eclipse and the comparison photographs to meet the same standard of representational accuracy would be to rule out evidence needed to support (E). Furthermore, the truth values of the observation sentences had better not be decided solely on the basis of whether the measurements of distances between their star images meet some general standard of accuracy specified independently of the particular investigation in question. In his textbook on error analysis, John Taylor points out that even though measurements can be too inaccurate to serve their purposes

> ...it is not necessary that the uncertainties [i.e. levels of error] be extremely small....[t]his...is typical of many scientific measurements, where uncertainties have to be reasonably small (perhaps a few percent of the measured value), but where extreme precision is often quite unnecessary.[32]

We maintain that what counts as a "reasonably small" level of error depends upon the nature of the phenomenon under investigation, the methods used to investigate it, and the alternative phenomena claims under consideration. Since these vary from case to case no single level of accuracy can distinguish between acceptable and unacceptable measurements for every case. Thus Priestley's nitric oxide test tolerates considerably more measurement error than did the starlight bending investigations. This means that in order to decide whether or not to treat observation sentences representing Eddington's photographs and measurements as true, the IRS epistemologist would have to know enough

---

[31] We mean measurable discrepancies not accounted for by changes in the position of the earth, differences in the location of the eclipse and comparison equipment, etc.

[32] Taylor (1982), p. 6.

about local details peculiar to their production and interpretation to find out what levels of error would be acceptable.

Suppose that one responds to this difficulty by stipulating that whatever observation sentences are used to represent photographs are to be called true if the photographs constitute good evidence and false if they do not. This means that the truth values of the observation sentences will depend, for example, upon whether the investigators could rule out or correct for the influences of such factors as mechanical changes in the equipment, parallax, sources of measurement error, etc., as far as necessary to allow them to discriminate correctly between (E), (N) and (NS). We submit that this stipulation is completely unilluminating. The notion of truth as applied to an observation sentence is now unconnected with any notion of representational correctness or accuracy (i.e., it is unclear what such sentences are supposed to represent or correspond to when they are true). Marking an observation sentence as true is now just a way of saying that the data associated with the sentence possess various other features that allow them to play a role in reliable discrimination. It is better to focus directly on the data and the processes that generate them and to drop the role of an observation sentence as an unnecessary intermediary.

VI

Recall that an important part of the motivation for the development of IRS was the question of what objective factors do or should determine a scientist's decision about whether a given body of evidence warrants the acceptance of a hypothesis. We have suggested that the evidential value of data depends upon complex and multifarious causal connections between the data, the phenomenon of interest, and a host of other factors. But it does not follow from this that scientists typically do (or even can) know much about the fine details of the relevant causal mechanisms. Quite the contrary, as we have argued elsewhere, scientists can seldom if ever give, and are seldom if ever required to give, detailed, systematic causal accounts of the production of a particular bit of data or its interaction with the human perceptual system and with devices (like the measuring equipment used by the starlight investigators) involved in its interpretation.[33] But even though it does not involve systematic causal explanation, we believe that a kind of causal reasoning is essential to the use of data to investigate phenomena. This reasoning focuses upon what we have called general and local reliability. The remainder of this paper discusses some features of this sort of reasoning, and argues that its objectivity does not depend upon, and is not well explained in terms of the highly general, content independent, formal criteria sought by the IRS.

---

[33] Bogen and Woodward (1988). For an excellent illustration of this, see Hacking (1983), p. 209.

VII

We turn first to a more detailed characterization of what we mean by general reliability. As we have already suggested, general reliability is a matter of long-run error characteristics. A detection process is generally reliable, when used in connection with a body of data, if it has a satisfactorily high probability of outputting, under repeated use, correct discriminations among a set of competing phenomenon-claims and a satisfactorily low probability of outputting incorrect discriminations. What matters is thus that the process discriminates correctly among a set of relevant alternatives, not that it discriminates correctly among all logically possible alternatives. Whether or not a detection process is generally reliable is always an empirical matter, having to do with the causal characteristics of the detection process and its typical circumstances of use, rather than with any formal relationship between the data that figure in such a process and the phenomenon-claims for which they are evidence. The notion of general reliability thus has application in those contexts in which we can provide a non-trivial characterization of what it is to repeat a process of data production and interpretation (we shall call this a detection process, for brevity) and where this process possesses fairly stable, determinate error characteristics under repetition that are susceptible of empirical investigation. As we shall see in section VIII, these conditions are met in many, but by no means all the contexts in which data are used to assess claims about phenomena. Where these conditions are not met, we must assess evidential support in terms of a distinct notion of reliability, which we call local reliability.

Here is an example illustrating what we have in mind by general reliability.[34] Traditionally paleoanthropologists have relied on fossil evidence to infer relationships among human beings and other primates. The 1960s witnessed the emergence of an entirely distinct biochemical method for making such inferences, which involved comparing proteins and nucleic acids from living species. This method rests on the assumption that the rate of mutation in proteins is regular or clocklike; with this assumption one can infer that the greater the difference in protein structure among species, the longer the time they have been separated into distinct species. Molecular phylogeny (as such techniques came to be called) initially suggested conclusions strikingly at variance with the more traditional, generally accepted conclusions based on fossil evidence. For example, while fossil evidence suggested an early divergence between hominids and other primates, molecular techniques suggested a much later date of divergence – that hominids appeared much later than previously thought. Thus while paleoanthropologists classified the important prehistoric primate Ramapithicus as an early hominid on the basis of its fossil remains, the molecular evidence seemed to suggest that Ramapithicus could not

[34] Details of this example are largely taken from Lewin (1987).

be a hominid. Similarly, fossil and morphological data seemed to suggest that chimpanzees and gorillas were more closely related to each other than to humans, while molecular data suggested that humans and chimpanzees were more closely related.

The initial reaction of most paleoanthropologists to these new claims was that the biochemical methods were unreliable, because they produced results at variance with what the fossils suggested. It was suggested that because the apparent rates of separation derived from molecular evidence were more recent than those derived from the fossil record, this showed that the molecular clock was not steady and that there had been a slow down in the rate of change in protein structure among hominids. This debate was largely resolved in favor of the superior reliability of molecular methods. The invention of more powerful molecular techniques based on DNA hybridization, supported by convincing statistical arguments that the rate of mutation was indeed clocklike, largely corroborated the results of earlier molecular methods. The discovery of additional fossil evidence undermined the hominid status of Ramapithicus and supported the claim of a late divergence between hominids and other primates.

This example illustrates what we have in mind when we ask whether a measurement or detection technique is generally reliable. We can think of various methods for inferring family trees from differences in protein structure and methods for inferring such relationships from fossil evidence as distinct measurement or detection techniques. Any particular molecular method is assumed to have fairly stable, determinate error characteristics which depend upon empirical features of the method: if the method is reliable it will generally yield roughly correct conclusions about family relationship and dates of divergence; if it is unreliable it will not. Clearly the general reliability of the molecular method will depend crucially on whether it is really true that the molecular clock is regular.

Similarly, the reliability of the method associated with the use of fossil evidence also depends upon a number of empirical considerations – among them the ability of human beings to detect overall patterns of similarity based on visual appearance that correlate with genetic relationships. What the partisans of fossils and of molecular methods disagree about is the reliability of the methods they favor, in just the sense of reliability as good error characteristics described above. Part of what paleoanthropologists learned as they became convinced of the superior reliability of molecular methods, was that methods based on similarity of appearance were often less reliable than they had previously thought, in part because judgements of similarity can be heavily influenced by prior expectations and can lead the investigator to think that she sees features in the fossil evidence that are simply not there.[35]

---

[35] See especially Lewin (1987), p. 122ff.

Issues of this sort about general reliability – about the long-run error characteristics of a technique or method under repeated applications – play a central role in many areas of scientific investigation. Whenever a new instrument or detection device is introduced, investigators will wish to know about its general reliability – whether it works in such a way as to yield correct discriminations with some reasonable probability of success, whether it can be relied upon as a source of information in some particular area of application. Thus Galileo's contemporaries were interested not just in whether his telescopic observations of the rough and irregular surface of the moon were correct, but with the general reliability of his telescope – with whether its causal characteristics were such that it could be used to make certain kinds of discrimination in astronomical applications with some reasonable probability of correctness or with whether instead what observers seemed to see through the telescope were artifacts, produced by imperfections in the lenses or some such source.

Similarly in many contexts in which human perceivers play an important role in science one can ask about their general reliability at various perceptual detection tasks, where this has to do with the probability or frequency with which perceivers make the relevant perceptual discriminations correctly, under repeated trials. Determinations of personal error rates in observational sciences like astronomy make use of this understanding of reliability.[36] Similarly one can ask whether an automated data reduction procedure which sorts through batches of photographs selecting those which satisfy some preselected criterion is operating reliably, where this has to do with whether or not it is in fact classifying the photographs according to the indicated criterion with a low error rate.

There are several general features of the above examples which are worth underscoring. Let us note to begin with that the question of whether a method, technique or detection device and the data it produces are reliable always depends very much on the specific features of the method, technique or instrument in question. It is these highly specific empirical facts about the general reliability of particular methods of data production and interpretation and not the formal relationships emphasized by IRS that are relevant to determining whether or not data are good evidence for various claims about phenomena. For example, it is the reliability characteristics of Galileo's telescope that insure the evidential relevance of the images that it produces to the astronomical objects he wishes to detect and it is the reliability characteristics of DNA hybridization that insure the evidential relevance of the biochemical data it produces to the reconstruction of relationships between species.

How is the general reliability of an instrument or detection technique ascertained? We (and others) have discussed this issue at some length elsewhere and readers are referred to this discussion for a more detailed treatment.[37]

---

[36] For additional discussion, see Bogen and Woodward (1992).

[37] See Bogen and Woodward (1988) and Woodward (1989). Although, on our view, it is always a matter of empirical fact whether or not a detection process is generally reliable, we want to

A wide variety of different kinds of considerations having to do, for example, with the observed effects of various manipulations and interventions into the detection process, with replicability, and with the use of various calibration techniques play an important role. One point that we especially wish to emphasize, and which we will make use of below, is that assessing the general reliability of an instrument or detection technique does not require that one possess a general theory that systematically explains the operation of the instrument or technique or why it is generally reliable. There are many cases in the history of science involving instruments and detection techniques that investigators reasonably believed to be generally reliable in various standard uses even though those investigators did not possess a general explanatory theory of the operation of these instruments and techniques. Thus it was reasonable of Galileo and his contemporaries to believe that his telescope was generally reliable in many of its applications, even though Galileo lacked an optical theory that explained its workings; it is reasonable to believe that the human visual system can reliably make various perceptual discriminations in specified circumstances even though our understanding of the operation of the visual system is still rudimentary; it may be reasonable to believe that a certain staining technique reliably stains certain cells and doesn't produce artifacts even though one doesn't understand the chemistry of the staining process, and so on.

We may contrast the picture we have been advocating, according to which evidential relevance is carried by the reliability characteristics of highly specific processes of data production and interpretation, with the conception of evidential relevance which is implicit in IRS. According to that conception, the relevance of evidence to hypotheses is a matter of observation sentences standing in various highly general, structural or inferential relations to those hypotheses, relationships which, according to IRS, are exemplified in many different areas of scientific investigation. Thus the idea is that the evidential relevance of biochemical data to species relationships or the evidential relevance of the images produced by Galileo's telescope to various astronomical hypotheses is a matter of the obtaining of some appropriate formal relationship between sentences representing these data, the hypotheses in question and perhaps appropriate background or auxiliary assumptions. On the contrasting picture we have defended, evidential relevance is not a matter of any such formal relationship, but is instead a matter of empirical fact – a matter of there

---

emphasize that there is rarely if ever an algorithm or mechanical procedure for deciding this. Instead it is typically the case that a variety of heterogeneous considerations are relevant, and building a case for general reliability or unreliability is a matter of building a consensus that most of these considerations, or the most compelling among them, support one conclusion rather than another. As writers like Peter Galison (1987) have emphasized, reaching such a conclusion may involve an irreducible element of judgement on the part of experimental investigators about which sources of error need to be taken seriously, about which possibilities are physically realistic, or plausible and so forth. Similar remarks apply to conclusions about local reliability. (Cf. n. 42).

existing empirical relationships or correlations between data and phenomena which permit us to use the data to discriminate among competing claims about phenomena according to procedures that have good general error characteristics. Evidential relevance thus derives from an enormous variety of highly domain specific facts about the error characteristics of various quite heterogeneous detection and measurement processes, rather than from the highly general, domain-independent formal relationships emphasized in IRS accounts.

Our alternative conception seems to us to have several advantages that are not shared by IRS accounts. First, we have already noted that a great deal of data does not have an obvious sentential representation and that, even when such representations are available, they need not be true or exactly representationally accurate for data to play an evidential role. Our account helps to make sense of these facts. There is nothing in the notion of general reliability that requires that data be sentential in structure, or have a natural sentential representation or have semantic characteristics like truth or exact representational accuracy. Data can figure in a generally reliable detection process, and features of data can be systematically correlated with the correctness or incorrectness of different claims about phenomena without the data being true or even sententially representable. For example, when a pathologist looks at an x-ray photograph and produces a diagnosis, or when a geologist looks at a rock and provides an identification of its type, all that we require, in order for these claims to be credible or evidentially well-supported, is that the relevant processes of perceptual detection and identification be generally reliable in the sense of having good error characteristics and that we have some evidence that this is the case. It isn't necessary that we be able to provide sentential representations of what these investigators perceive or to exhibit their conclusions as the result of the operation of general IRS-style inductive rules on sentential representations of what they see. Similarly, in the case of the Priestley/Lavoisier example, the characteristics of Priestley's detection procedure may very well be such that it can be used to reliably discriminate between ordinary air and oxygen on the basis of volume measurements, in the sense that repeated uses of the procedure will result in correct discriminations with high probability, even though the volume measurements on which the discrimination is based are inaccurate, noisy and in fact false if taken as reports of the actual volume decrease.

There is a second reason to focus on reliability in preference to IRS-style confirmation relations. According to the IRS, evidence *e* provides epistemic support for a theoretical claim when the observation sentence, *o*, which corresponds to the evidence stands in the right sort of formal relationship to the hypothesis sentence, *h*, which represents the theoretical claim. Our worries so far have centered around the difficulties of finding a true observation sentence *o* which faithfully represents the evidential significance of *e*, and a hypothesis

sentence *h* which faithfully represents the content of the theoretical claim. But quite apart from these difficulties there is a perennial internal puzzle for IRS accounts. Given that within these accounts *o* does not, even in conjunction with background information, entail *h*, why should we suppose that there is any connection between *o*'s being true and *o* and *h* instantiating the formal relationships specified in these accounts and *h*'s being true or having a high probability of truth or possessing some other feature associated with grounds for belief? For example, even if a true observation sentence representing Priestley's data actually did entail a positive instance of a hypothesis sentence representing the claim that a certain sort of gas is not ordinary air, why would that make the latter claim belief-worthy? We think that it is very hard to see what the justification of a non-deductive IRS-style method or criterion of evidential support could possibly consist in except the provision of grounds that the method or criterion has good (general) reliability or error characteristics under repeated use. That is, it is hard to see why we should believe that the truth of the observation sentence *o* together with the fact that the relationship between *o* and hypothesis *h* satisfies the pattern recommended by, for example, hypothetico-deductivism or bootstrapping provides a reason for belief in *h* if it were not true that cases in which such patterns are instantiated turn out, with some reasonable probability, to be cases in which *h* is true or were it not at least true that cases in which such patterns are instantiated turn out more frequently to be cases in which *h* is true than cases in which such patterns are not instantiated.[38]

However, it seems very unlikely that any of the IRS-style accounts we have considered can be given such a reliabilist justification. IRS accounts are, as we have seen, subject matter and context-independent; they are meant to supply universal criteria of evidential support. But it is all too easy to find, for any IRS account, not just hypothetical, but actual cases in which true observation sentences stand in the recommended relationship to hypothesis *h* and yet in which *h* is false: cases in which positive instances instantiate a hypothesis and yet the hypothesis is false, cases in which true observation sentences are deduced from a hypothesis and yet it is false, and so forth. Whether accepting *h* when it stands in the relationship to *o* described in one's favorite IRS schema and *o* is true will lead one to accept true hypotheses some significant fraction of the time will depend entirely on the empirical details of the particular cases to which the schema in question is applied. But this is to say that the various IRS

[38] For a general argument in support of this conclusion see Friedman (1979). One can think of Larry Laudan's recent naturalizing program in philosophy of science which advocates the testing of various philosophical theses about scientific change and theory confirmation against empirical evidence provided by the history of science as (among other things) an attempt to carry out an empirical investigation of the error or reliability characteristics of the various IRS confirmation schemas (Donovan *et al.*, 1988). We agree with Laudan that vindicating the various IRS models would require information about long-run error characteristics of the sort for which he is looking. But for reasons described in the next paragraph in the text, we are much more pessimistic than Laudan and his collaborators about the possibility of obtaining such information.

schemas we have been considering when taken as methods for forming beliefs or accepting hypotheses either have no determinate error characteristics at all when considered in the abstract (their error characteristics vary wildly, depending on the details of the particular cases to which they are applied) or at least no error characteristics that are knowable by us. Indeed, the fact that the various IRS accounts we have been considering cannot be given a satisfying reliabilist justification is tacitly conceded by their proponents, who usually do not even try to provide such a justification.[39]

By contrast, there is no corresponding problem with the notion of general reliability as applied to particular instruments or detection processes. Such instruments and processes often do have determinate error characteristics, about which we can obtain empirical evidence. Unlike the H-D method or the method associated with bootstrapping, the reliability of a telescope or a radioactive dating technique is exactly the sort of thing we know how to investigate empirically and regarding which we can obtain convincing evidence. There is no puzzle corresponding to that raised above in connection with IRS accounts about what it means to say that a dating technique has a high probability of yielding correct conclusions about the ages of certain fossils or about why, given that we have applied a reliable dating technique and have obtained a certain result, we have good *prima facie* grounds for believing that result. In short, it is the use of specific instruments, detection devices, measurement and observational techniques, rather than IRS-style inductive patterns, that are appropriate candidates for justification in terms of the idea of general reliability. Reflection on a reliabilist conception of justification thus reinforces our conclusion that the relevance of evidence to hypothesis is not a matter of formal, IRS-style inferential relations, but rather derives from highly specific facts about the error characteristics of various detection processes and instruments.

---

[39] Typical attempts to argue for particular IRS models appeal instead to (a) alleged paradoxes, and inadequacies associated with alternative IRS approaches, (b) various supposed intuitions about evidential support, and (c) famous examples of successful science that are alleged to conform to the model in question. (Cf. Glymour (1980).) But (a) is compatible with and perhaps even supports skepticism about all IRS accounts of evidence, and with respect to (b), it is uncontroversial that intuitions about inductive support frequently lead one astray. Finally, from a reliabilist perspective (c) is quite unconvincing. Instead, what needs to be shown is that scientists systematically succeed in a variety of cases because they accept hypotheses in accord with the recommendations of the IRS account one favors. That is, what we need to know is not just that there are episodes in the history of science in which hypotheses stand in the relationship to true observation sentences described by, say, a bootstrap methodology and that these hypotheses turn out to be true or nearly so, but what the performance of a bootstrap methodology would be, on a wide variety of different kinds of evidence, in discriminating true hypotheses from false hypotheses – both what this performance is absolutely and how it compares with alternative methods one might adopt. (As we understand it, this is Glymour's present view as well.)

VIII

In addition to the question of whether some type of detection process or instrument is generally reliable in the repeatable error characteristics sense described above, scientists also are interested in whether the use of the process on some particular occasion, in a particular detection task, is reliable – with whether the data produced on that particular occasion are good evidence for some phenomenon of interest. This is a matter of *local* reliability. While in those cases in which a detection process has repeatable error characteristics, information about its general reliability is always evidentially relevant, there are many cases in which the evidential import of data cannot be assessed just in terms of general reliability. For example, even if I know that some radioactive dating technique is generally reliable when applied to fossils, this still leaves open the question of whether the date assigned to some particular fossil by the use of the technique is correct: it might be that this particular fossil is contaminated in a way that gives us mistaken data, or that the equipment I am using has misfunctioned on this particular occasion of use. That the dating process is generally reliable doesn't preclude these possibilities.

Some philosophers with a generalist turn of mind will find it tempting to try to reduce local reliability to general reliability: it will be said that if the data obtained from a particular fossil is mistaken because of the presence of a contaminant, then if that very detection process is repeated (with the contaminant present and so forth) on other occasions, it will have unfavorable error characteristics, and this is what grounds our judgement of reliability or evidential import in the particular case. As long as we take care to specify the relevant detection processes finely enough, all judgements about reliability in particular cases can be explicated in terms of the idea of repeated error characteristics. Our response is not that this is necessarily wrong, but that it is thoroughly unilluminating at least when understood as an account of how judgements of local reliability are arrived at and justified. As we shall see below, many judgements of local reliability turn on considerations that are particular or idiosyncratic to the individual case at hand. Often scientists are either unable to describe in a non-trivial way what it is to repeat the measurement or detection process that results in some particular body of data or lack (and cannot get) information about its long-run error characteristics. It is not at all clear to us that whenever a detection process is used on some particular occasion, and a judgement about its local reliability is reached on the basis of various considerations, there must be some description of the process, considerations, and judgements involved that exhibits them as repeatable. But even if this is the case, this description and the relevant error characteristics of the process when repeated often will be unknown to the individual investigator – this information is not what the investigator appeals to in reaching his judgement about local reliability or in defending his judgement.

What then are the considerations which ground judgements of local reliability and how should we understand what it is that we are trying to do when we make such judgements? While the relevant considerations are, as we shall see, highly heterogeneous, we think that they very often have a common point or pattern, which we will now try to describe. Put baldly, our idea is that judgements of local reliability are a species of singular causal inference in which one tries to show that the phenomenon of interest causes the data by means of an eliminativist strategy – by ruling out other possible causes of the data.[40] When one makes a judgement of local reliability one wants to ascertain on the basis of some body of data whether some phenomenon of interest is present or has certain features. One tries to do this by showing that the detection process and data are such that the data must have been caused by the phenomenon in question (or by a phenomenon with the features in question) – that all other relevant candidates for causes of the data can be ruled out. Since something must have caused the data, we settle on the phenomenon of interest as the only remaining possibility. For example, in the fossil dating example above, one wants to exclude (among other things) the possibility that one's data – presumably some measure of radioactive decay rate, such as counts with a Geiger counter – were caused by (or result in part from a causal contribution due to) the presence of the contaminant. Similarly, as we have already noted, showing that some particular bubble chamber photograph was evidence for the existence of neutral currents in the CERN experiments of 1973 requires ruling out the possibility that the particular photograph might have been due instead to some alternative cause, such as a high energy neutron, that can mimic many of the effects of neutral currents. The underlying idea of this strategy is nicely described by Allan Franklin in his recent book *Experiments, Right or Wrong* (1990). Franklin approvingly quotes Sherlock Holmes's remark to Watson, "How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?" and then adds, "If we can eliminate all possible sources of error and alternative explanations, then we are left with a valid experimental result" (1990, p. 109).

Here is a more extended example designed to illustrate what is involved in local reliability and the role of the eliminative strategy described above.[41] In experiments conducted in the late 1960s, Joseph Weber, an experimentalist at the University of Maryland, claimed to have successfully detected the

---

[40] As with judgements about general reliability, we do not mean to suggest that there is some single method or algorithm to be employed in this ruling out of alternatives. For example, ruling out an alternative may involve establishing an observational claim that is logically inconsistent with the alternative (Popperian falsification), but might take other forms as well; for example, it may be a matter of finding evidence that renders the alternative unlikely or implausible or of finding evidence that the alternative should but is not able to explain.

[41] The account that follows draws heavily on Collins (1975) and Collins (1981). Other accessible discussions of Weber's experiment on which we have relied include Davis (1980), esp. pp. 102-117, and Will (1986).

phenomenon of gravitational radiation. The production of gravity waves by massive moving bodies is predicted (and explained) by general relativity. However, gravitational radiation is so weakly coupled to matter that detection of such radiation by us is extremely difficult.

Weber's apparatus initially consisted of a large metal bar which was designed to vibrate at the characteristic frequency of gravitational radiation emitted by relatively large scale cosmological events. The central problem of experimental design was that to detect gravitational radiation one had to be able to control or correct for other potential disturbances due to electromagnetic, thermal, and acoustic sources. In part, this was attempted by physical insulation of the bar, but this could not eliminate all possible sources of disturbance; for example, as long as the bar is above absolute zero, thermal motion of the atoms in the bar will induce random vibrations in it. One of the ways Weber attempted to deal with this difficulty was through the use of a second detector which was separated from his original detector by a large spatial distance – the idea being that genuine gravitational radiation, which would be cosmological in origin, should register simultaneously on both detectors while other sorts of background events which were local in origin would be less likely to do this. Nonetheless, it was recognized that some coincident disturbances will occur in the two detectors just by chance. To deal with this possibility, various complex statistical arguments and other kinds of checks were used to attempt to show that it was unlikely that all of the coincident disturbances could arise in this way.

Weber also relied on facts about the causal characteristics of the signal – the gravitational radiation he was trying to detect. The detectors used by Weber were most sensitive to gravitational radiation when the direction of propagation of given radiation was perpendicular to the axes of detectors. Thus if the waves were coming from a fixed direction in space (as would be plausible if they were due to some astronomical event), they should vary regularly in intensity with the period of revolution of the earth. Moreover, any periodic variations due to human activity should exhibit the regular twenty-four hour variation of the solar day. By contrast, the pattern of change due to an astronomical source would be expected to be in accordance with the sidereal day which reflects the revolution of the earth around the sun, as well as its rotation about its axis, and is slightly shorter than the solar day. When Weber initially appeared to find a significant correlation with sidereal, but not solar, time in the vibrations he was detecting, this was taken by many other scientists to be important evidence that the source of the vibrations was not local or terrestrial, but instead due to some astronomical event.

Weber claimed to have detected the existence of gravitational radiation from 1969 on, but for a variety of reasons his claims are now almost universally doubted. In what follows, we concentrate on what is involved in Weber's claim that his detection procedure was locally reliable and how he attempted to

establish that claim. As we see it, what Weber was interested in establishing was a singular causal claim: he wanted to show that at least some of the vibrations and disturbances his data recorded were due to gravitational radiation (the phenomenon he was trying to detect) and (hence) that such radiation existed. The problem he faced was that a number of other possible causes or factors besides gravitational radiation might in principle have caused his data. Unless Weber could rule out, or render implausible or unlikely, the possibility that these other factors might have caused the disturbances, he would not be justified in concluding that the disturbances are due to the presence of gravitational radiation. The various experimental strategies and arguments described above (physical isolation of the bar, use of second detector, and so forth) are an attempt to do just this – to make it implausible that the vibrations in his detector could have been caused by anything but gravitational radiation. For example, in the case of the sidereal correlation the underlying argument is that the presence of this pattern or signature in the data is so distinctive that it could only have been produced by gravitational radiation rather than by some other source.

We will not attempt to describe in detail the process by which Weber's claims of successful detection came to be criticized and eventually disbelieved. Nonetheless it is worth noting that we can see the underlying point of these criticisms as showing that Weber's experiment fails to conform to the eliminative pattern under discussion – what the critics show is that Weber has not convincingly ruled out the possibility that his data were due to other causes besides gravitational radiation. Thus, for example, the statistical techniques that Weber used turned out to be problematic – indeed, an inadvertent natural experiment appeared to show that the techniques lacked general reliability in the sense described above. (Weber's statistical techniques detected evidence for gravitational radiation in data provided by another group which, because of a misunderstanding on Weber's part about synchronization, should have been reported as containing pure noise.) Because of this, Weber could no longer claim to have convincingly eliminated the possibility that all of the disturbances he was seeing in both detections were due to the chance coincidence of local causes.

Secondly, as Weber continued his experiment and did further analysis of his data, he was forced to retract his claim of sidereal correlation. Finally, and perhaps most fundamentally, a number of other experiments, using similar and more sensitive apparatus, failed to replicate Weber's results. Here the argument is that if in fact gravitational radiation was playing a causal role in the production of Weber's data such radiation ought to interact causally with other similar devices; conversely, failure to detect such radiation with a similar apparatus, while it does not tell us which alternative cause produced Weber's data, does undermine the claim that it was due to gravitational radiation.

Much of what we have said about the advantages of the notion of general reliability vis-à-vis IRS-style accounts holds as well for local reliability. When

we make a judgement of local reliability about certain data – when we conclude, for example, that some set of vibrations in Weber's apparatus were or were not evidence for the existence of gravitational radiation – what needs to be established is not whether there obtains some appropriate formal or logical relationship of the sort IRS models attempt to capture, but rather whether there is an appropriate causal relationship leading from the phenomenon to the data. Just as with general reliability, the causal relationships needed for data to count as locally reliable evidence for some phenomenon can hold even if data lack a natural sentential representation that stands in the right formal relationship to the phenomenon-claim in question.

Conversely, a sentential representation of the data can stand in what (according to some IRS accounts of confirmation) is the right formal relationship to a hypothesis and yet nonetheless fail to evidentially support it. Weber's experiment also illustrates this point: Weber obtained data which (or so he was prepared to argue) were just what would be expected if general relativity were true (and gravitational radiation existed). On at least some natural ways of representing data by means of observation sentences, these sentences stand in just the formal relationships to general relativity which according to H-D and positive instance accounts, are necessary for confirmation. Nonetheless this consideration does *not* show that Weber's data were reliable evidence for the existence of gravitational radiation. To show this Weber must show that his data were produced by a causal process in which gravitational radiation figures. This is exactly what he tries, and fails, to do. The causally driven strategies and arguments described above would make little sense if all Weber needed to show was the existence of some appropriate IRS-style formal relationship between a true sentential representation of his data and the claim that gravitational radiation exists. Similarly, as we have already had occasion to note, merely producing bubble chamber photographs that have just the characteristic patterns that would be expected if neutral currents were present – producing data which conform to this hypothesis or which have some description which is derivable from the hypothesis – is not by itself good evidence that neutral currents are present. To do this one must rule out the possibility that this data was caused by anything but neutral currents. And as we have noted, this involves talking about the causal process that has produced the data – a consideration which is omitted in most IRS accounts.

As we have also argued, a similar point holds in connection with the Eddington solar eclipse expedition. What Eddington needs to show is that the apparent deflection of starlight indicated by the photographic plates is due to the causal influence of the sun's gravitational field, as described by general relativity, rather than to more local sources, such as changes in the plates due to variations in temperature. Once we understand Eddington's reasoning as reasoning to the existence of a cause in accordance with an eliminative strategy, various features of that reasoning that seem puzzling on IRS treatments – that it

is not obvious how to represent all of the evidentially relevant features of the photographs in terms of true observation sentences and auxiliaries and that the values calculated from the photographs don't exactly coincide with (E) but are nonetheless taken to support (E) – fall naturally into place.

IX

There is a common element to a number of the difficulties with IRS models that we have discussed that deserves explicit emphasis. It is an immediate consequence of our notions of general and local reliability that the processes that produce or generate data are crucial to its evidential status. Moreover, it is often hard to see how to represent the evidential relevance of such processes in an illuminating way within IRS-style accounts. And in fact the most prominent IRS models simply neglect this element of evidential assessment. The tendency within IRS models is to assume, as a point of departure, that one has a body of evidence, that it is unproblematic how to represent it sententially, and to then try to capture its evidential relevance to some hypothesis by focusing on the formal or structural relationship of its sentential representation to that hypothesis. But if the processes that generated this evidence make a crucial difference to its evidential significance, we can't as IRS approaches assume, simply detach the evidence from the processes which generated it, and use a sentential representation of it as a premise in an IRS-style inductive inference.

To make this point vivid, consider (P) a collection of photographs which qua photographs are indistinguishable from those that in fact constituted evidence for the existence of neutral current interactions in the CERN experiments of 1973. Are the photographs in P also evidence for the existence of neutral currents? Although many philosophers (influenced by IRS models of confirmation) will hold that the answer to this question is obviously yes, our claim is that on the basis of the above information one simply doesn't know – one doesn't know whether the photographs are evidence for neutral currents until one knows something about the processes by which they are generated. Suppose that the process by which the photographs were produced failed to adequately control for high energy neutrons. Then our claim is that photographs are not reliable evidence for the existence of neutral currents, even if the photographs themselves look no different from those that were produced by experiments (like the CERN experiment) in which there was adequate control for the neutron background. It is thus a consequence of our discussion of general and local reliability that the evidential significance of the same body of data will vary, depending upon what it is reasonable to believe about how it was produced.

We think that the tendency to neglect the relevance of the data-generating processes explains, at least in large measure, the familiar paradoxes which face

IRS accounts. Consider the raven paradox, briefly introduced in section IV above. Given our discussion so far it will come as no surprise to learn that we think the culprit in this case is the positive instance criterion itself. Our view is that one just can't say whether a positive instance of a hypothesis constitutes evidence for it, without knowing about the procedure by which the positive instance was produced or generated. A possibility originally introduced by Paul Horwich (1982) makes this point in a very striking way: suppose that you are told that a large number of ravens have been collected, and that they have all turned out to be black. You may be tempted to suppose that such observations support the hypothesis that (h_1) all ravens are black. Suppose, however, you then learn how this evidence has been produced: a machine of special design which seizes all and only black objects and stores them in a vast bin has been employed, and all of our observed ravens have come from this bin. In the bin, we find, unsurprisingly, in addition to black shoes, old tires and pieces of coal, a number of black ravens and no non-black ravens.

Recall that our interest in data is in using it to discriminate among competing phenomenon-claims. Similarly, when we investigate the hypothesis that all ravens are black, our interest is in obtaining evidence that differentially supports this hypothesis against other natural competitors. That is, our interest is in whether there is evidence that provides some basis for preferring or accepting this hypothesis in contrast to such natural competitors as the hypothesis that ravens come in many different colors, including black. It is clear that the black ravens produced by Horwich's machine do not differentially support the hypothesis that all ravens are black or provide grounds for accepting it rather than such competitors. The reason is obvious: the character of the evidence gathering or data-generating procedure is such that it could not possibly have discovered any evidence which is contrary to the hypothesis that all ravens are black, or which discriminates in favor of a competitor to this hypothesis, even if such evidence exists. The observed black ravens are positive instances of the hypothesis that all ravens are black, but they do not support the hypothesis in the sense of discriminating in favor of it against natural competitors because of the way in which those observations have been produced or generated. If observations of a very large number of black ravens had been produced in some other way – e.g., by a random sampling process, which had an equal probability of selecting any raven (black or non-black) or by some other process which was such that there was some reason to think that the evidence it generated was representative of the entire population of ravens – then we would be entitled to regard such observations as providing evidence that favors the hypothesis under discussion. But in the absence of a reason to think that the observations have been generated by some such process that makes for reliability, the mere accumulation of observations of black ravens provides no reason for accepting the hypothesis that all ravens are black in contrast to its natural competitors.

Similar considerations apply to the question of whether the observation of non-black, non-ravens supports the hypothesis that (h_2), "All non-black things are non-ravens." As a point of departure, let us note that it is less clear than it is in the case of (h_1) what the "natural" serious alternatives to (h_2) are. The hypothesis (h_3) that "All non-black things are ravens" is a competitor to (h_2) – it is inconsistent with (h_2) on the supposition that there is at least one non-black thing – but not a serious competitor since every investigator will have great confidence that it is false prior to beginning an investigation of (h_2). Someone who is uncertain whether (h_2) is true will not take seriously the possibility that (h_3) is true instead and for this reason evidence that merely discriminates between (h_2) and (h_3) but not between (h_2) and its more plausible alternatives will not be regarded as supporting (h_2). Thus while the observation of a white shoe does indeed discriminate between (h_2) and (h_3) this fact by itself does not show that the observation supports (h_2). Presumably the best candidates for serious specific alternatives to (h_2) are various hypotheses specifying the conditions (e.g., snowy regions) under which non-black ravens will occur. But given any plausible alternative hypothesis about the conditions under which a non-black raven will occur, the observation of a white shoe or a red pencil does nothing to effectively discriminate between (h_2) and this alternative. For example, these observations do nothing to discriminate between (h_2) and the alternative hypotheses that there are white ravens in snowy regions. As far as these alternatives go, then, there is no good reason to think of an observation of a white shoe as confirming (h_2).

There are other possible alternatives to (h_2) that one might consider. For example, there are various hypotheses, (h_p), specifying that the proportion of ravens among non-black things is some (presumably very small) positive number p for various values of p. There is also the generic, non-specific alternative to (h_2) which is simply its denial (h_4), "Some non-black things are ravens." For a variety of reasons these alternatives are less likely to be of scientific interest than the alternatives considered in the previous paragraph. But even if we put this consideration aside, there is an additional problem with the suggestion that the observation of a white shoe confirms (h_2) because it discriminates between (h_2) and one or more of these alternatives.

This has to do with the characteristics of the processes involved in the production of such observations. In the case of (h_1), "All ravens are black," we have some sense of what it would mean to sample randomly from the class of ravens or at least to sample a "representative" range of ravens (e.g., from different geographical locations or ecological niches) from this class. That is we have in this case some sense of what is required for the process that generates relevant observations to be unbiased or to have good reliability characteristics. If we observe enough ravens that are produced by such a process and all turn out to be black we may regard this evidence as undercutting not just those competitors to (h_1) that claim that all ravens are some uniform non-black color

but also those alternative hypotheses that claim that various proportions of ravens are non-black or the generic alternative hypothesis that some ravens are non-black. Relatedly, observations of non-black ravens produced by such a process might confirm some alternative hypothesis to (h_1) about the proportion of ravens that are non-black or the conditions under which we may expect to find them.

By contrast, nothing like this is true of (h_2). It is hard to understand even what it might mean to sample in a random or representative way from the class of non-black things and harder still to envision a physical process that would implement such a sampling procedure. It is also hard to see on what basis one might argue that a particular sample of non-black things was representative of the entire range of such things. As a result when we are presented with even a very generous collection of objects consisting of white shoes, red pencils and so on it is hard to see on what sort of basis one might determine whether the procedure by which this evidence was produced had the right sort of characteristics to enable us to reliably discriminate between (h_2) and either the alternatives (h_p) or (h_4), and hence hard to assess what its evidential significance is for (h_2). It is thus unsurprising that we intuitively judge the import of such evidence for (h_2) to be at best unclear and equivocal.

On our analysis, then, an important part of what generates the paradox is the mistaken assumption, characteristic of IRS approaches, that evidential support for a claim is just a matter of observation sentences standing in some appropriate structural or formal relationship to a hypothesis sentence (in this case the relationship captured by the positive instance criterion) independently of the processes which generate the evidence and independently of whether the evidence can be used to discriminate between the hypothesis and alternatives to it.

It might be thought that while extant IRS accounts have in fact neglected the relevance of those features of data-generating processes that we have sought to capture with our notions of general and local reliability, there is nothing in the logic of such accounts that requires this omission. Many IRS accounts assign an important role to auxiliary or background assumptions. Why can't partisans of IRS represent the evidential significance of processes of data generation by means of these assumptions?

We don't see how to do this in a way that respects the underlying aspirations of the IRS approach and avoids trivialization. The neglect of data generating processes in standard IRS accounts is not an accidental or easily correctable feature of such accounts. Consider those features of data generation captured by our notion of general reliability. What would the background assumptions designed to capture this notion within an IRS account look like? We have already argued that in order to know that an instrument or detection process is generally reliable, it is not necessary to possess a general theory that explains the operation of the instrument or the detection process. The background assumptions that are designed to capture the role of general reliability in

inferences from data to phenomena thus cannot be provided by general theories that explain the operation of instruments or detection processes. The information that grounds judgements of general reliability is, as we have seen, typically information from a variety of different sources – about the performance of the detection process in other situations in which it is known what results to expect, about the results of manipulating or interfering with the detection process in various ways, and so forth. While all of this information is relevant to reliability, no single piece of information of this sort is sufficient to guarantee reliability. Because this is the case and because the considerations which are relevant to reliability are so heterogeneous and so specific to the particular detection process we want to assess, it is not clear how to represent such information as a conventional background or auxiliary assumption or as a premise in an inductive inference conforming to some IRS pattern.

Of course we can represent the relevant background assumptions by means of the brute assertion that the instruments and detection processes with which we are working are generally reliable. Then we might represent the decision to accept phenomenon-claim P, on the basis of data D produced by detection process R as having something like the following structure: (1) If detection process R is generally reliable and produces data having features D, it follows that phenomenon-claim P will be true with high probability. (2) Detection process R is generally reliable and has produced data having features D; therefore (3) phenomenon-claim P is true with high probability (or alternatively (4) phenomenon-claim P is true). The problem with this, of course, is that the inference from data to phenomenon now no longer looks like an IRS-style inductive inference at all. The resulting argument is deductive if (3) is the conclusion. If (4) is the conclusion, the explicitly inductive step is trivial – a matter of adopting some rule of acceptance that allows one to accept highly probable claims as true. All of the real work is done by the highly specific subject-matter dependent background claim (2) in which general reliability is asserted. The original aspiration of the IRS approach, which was to represent the goodness of the inference as a matter of its conforming to some highly general, subject matter independent pattern of argument – with the subject matter independent pattern supplying, so to speak, the inductive component to the argument – has not been met.[42]

---

[42] Although we lack the space for a detailed discussion, we think that a similar conclusion holds in connection with judgements of local reliability. If one wished to represent formally the eliminative reasoning involved in establishing local reliability, then it is often most natural to represent it by means of the deductively valid argument pattern know as disjunctive syllogism: one begins with the premise that some disjunction is true, shows that all of the disjuncts save one are false, and concludes that the remaining disjunct is true. But, as in the case of the representation of the argument appealing to general reliability considered on p.** above, this formal representation of eliminative reasoning is obvious and trivial; the really interesting and difficult work that must be done in connection with assessing such arguments has to do with writing down and establishing the truth of their premises: has one really considered all the alternatives, does on really have good grounds for

Here is another way of putting this matter: someone who accepts (1) and (2) will find his beliefs about the truth of P significantly constrained, and constrained by empirical facts about evidence. Nonetheless the *kind* of constraint provided by (1) and (2) is very different from the kinds of non-deductive constraints on hypothesis choice sought by proponents of IRS models. Consider again the passage quoted from Hempel in section II. As that passage suggests, the aim of the IRS approach is to exhibit the grounds for belief in hypotheses like (3) or (4) in a way that avoids reference to "personal" or "subjective" factors and to subject-matter specific considerations. Instead the aim of the IRS approach is to exhibit the grounds for belief in (3) or (4) as resulting from the operation of some small number of general patterns of non-deductive argument or evidential support which recur across many different areas of inquiry. If (2) is a highly subject-matter specific claim about, say, the reliability of a carbon-14 dating procedure when applied to a certain kind of fossil or (even worse) a claim that asserts the reliability of a particular pathologist in correctly discriminating benign from malignant lung tumors when she looks at x-ray photographs, reference to "subject-matter specific" or "personal" considerations will not have been avoided. A satisfactory IRS analysis would begin instead with some sentential characterization of the data produced by the radioactive dating procedure or the data looked at by the pathologist, and then show us how this data characterization supports (3) or (4) by standing in some formally characterizable relationship to it that can be instantiated in many different areas of inquiry. That is, the evidential relevance of the data to (3) or (4) should be established or represented by the instantiation of some appropriate IRS pattern, not by a highly subject-matter specific hypothesis like (2). If our critique of IRS is correct, this is just what cannot be done.

As the passage quoted from Hempel makes clear, IRS accounts are driven in large measure by a desire to exhibit science as an objective, evidentially constrained enterprise. We fully agree with this picture of science. We think that in many scientific contexts, evidence has accumulated in such a way that only one hypothesis from some large class of competitors is a plausible candidate for belief or acceptance. Our disagreement with IRS accounts has to do with the nature or character of the evidential constraints that are operative in science, not with whether such constraints exist. According to IRS accounts these constraints derive from highly general, domain-independent, formally characterizable patterns of evidential support that appear in many different areas of scientific investigation. We reject this claim as well as Hempel's implied suggestion that either the way in which evidence constrains belief must be

---

considering all but one to be false? Answering such questions typically requires a great deal of subject-matter specific causal knowledge. Just as in the case of general reliability, the original IRS aspiration of finding a subject-matter independent pattern of inductive argument in which the formal features of the pattern do interesting, non-trivial work of a sort that might be studied by philosophers has not been met.

capturable within an IRS-style framework or else we must agree that there are no such constraints at all. On the contrasting picture we have sought to provide, the way in which evidence constrains belief should be understood instead in terms of non-formal subject-matter specific kinds of empirical considerations that we have sought to capture with our notions of general and local reliability. On our account, many well-known difficulties for IRS approaches – the various paradoxes of confirmation, and the problem of explaining the connection between a hypothesis standing in the formal relationships to an observation sentence emphasized in IRS accounts and its being true – are avoided. And many features of actual scientific practice that look opaque on IRS approaches – the evidential significance of data generating processes or the use of data that lacks a natural sentential representation, or that is noisy, inaccurate or subject to error – fall naturally into place.[43]

REFERENCES

Bogen, J. and Woodward, J. (1988). Saving the Phenomena. *The Philosophical Review*, 97, 303-52.

_____. (1992). Observations, Theories, and the Evolution of the Human Spirit. *Philosophy of Science*, 59, 590-611.

Braithwaite, R. (1953). *Scientific Explanation*. Cambridge: Cambridge University Press.

Collins, H. M. (1975). The Seven Sexes: A Study in the Sociology of a Phenomenon, or the Replication of Experiments in Physics. *Sociology*, 9, 205-24.

_____. (1981). Son of Seven Sexes: The Social Deconstruction of a Physical Phenomenon. *Social Studies of Science*, 11, 33-62.

Conant, J. B. (1957). The Overthrow of the Phlogiston Theory: The Chemical Revolution of 1775-1789. In J.B. Conant, L.K. Nash (eds.), *Harvard Case Histories in Experimental Science*, vol. 1. Cambridge, Mass.: Harvard University Press.

Davis, P. (1980). *The Search for Gravity Waves*. Cambridge: Cambridge University Press.

Donovan, A., Laudan, L., and Laudan, R. (1988). *Scrutinizing Science*. Dordrecht: Reidel.

---

[43] We have ignored Bayesian accounts of confirmation. We believe that in principle such accounts have the resources to deal with some although perhaps not all of the difficulties for IRS approaches described above. However, in practice the Bayesian treatments provided by philosophers often fall prey to these difficulties, perhaps because those who construct them commonly retain the sorts of expectations about evidence that characterize IRS style approaches. Thus while there seems no barrier in principle to incorporating information about the process by which data has been generated into a Bayesian analysis, in practice many Bayesians neglect or overlook the evidential relevance of such information – Bayesian criticisms of randomization in experimental design are one conspicuous expression of this neglect. For a recent illustration of how Bayesians can capture the evidential relevance of data generating processes in connection with the ravens paradox see Earman (1992); for a rather more typical illustration of a recent Bayesian analysis that fails to recognize the relevance of such considerations see the discussion of this paradox in Howson and Urbach (1989).

As another illustration of the relevance of the discussion in this paper to Bayesian approaches, consider that most Bayesian accounts require that all evidence have a natural representation by means of true sentences. These accounts thus must be modified or extended to deal with the fact that such a representation will not always exist. For a very interesting attempt to do just this, see Jeffrey (1989).

Earman, J. (1992*). Bayes or Bust?: A Critical Examination of Bayesian Confirmation Theory.* Cambridge, Mass.: MIT Press.
Earman, J. and Glymour, C. (1980). Relativity and Eclipses. In J.L. Heilbron (ed.), *Historical Studies in the Physical Sciences*, vol. 11, Part I.
Feyerabend, P. K. (1985). *Problems of Empiricism*. Cambridge: Cambridge University Press.
Franklin, A. (1990). *Experiment, Right or Wrong*. Cambridge: Cambridge University Press.
Friedman, M. (1979). Truth and Confirmation. *The Journal of Philosophy*, 76, 361-382.
Galison, P. (1987). *How Experiments End*. Chicago: University of Chicago Press.
Glymour, C. (1980). *Theory and Evidence*. Princeton: Princeton University Press.
Goldman, A. (1986). *Epistemology and Cognition.* Cambridge, MA: Harvard University Press.
Hacking, I. (1983). *Representing and Intervening*. Cambridge: Cambridge University Press.
Hempel, C. G. (1965) *Aspects of Scientific Explanation.* New York: The Free Press.
Howson, C. and Urbach, P. (1989). *Scientific Reasoning: The Bayesian Approach.* La Salle, Ill.: Open Court.
Humphreys, P. (1989). *The Chances of Explanation.* Princeton: Princeton University Press.
Jeffrey, R. (1989). Probabilizing Pathology. *Proceedings of the Aristotelian Society*, 89, 211-226.
Lavoisier, A. (1965). *Elements of Chemistry*. Translated by W. Creech. New York: Dover.
Lewin, R. (1987). *Bones of Contention.* New York: Simon and Schuster.
Mackie, J. L. (1963). The Paradox of Confirmation. *The British Journal for the Philosophy of Science*, 13, 265-277.
Merrill, G. H. (1979). Confirmation and Prediction. *Philosophy of Science*, 46, 98-117.
Miller, R. (1987). *Fact and Method*. Princeton: Princeton University Press.
Pais, A. (1982). *'Subtle is the Lord...': The Science and Life of Albert Einstein.* Oxford: Oxford University Press.
Popper, K. R. (1959). *The Logic of Scientific Discovery*. New York: Harper & Row.
Priestley, J. (1970). *Experiments and Observations on Different Kinds of Air, and Other Branches of Natural Philosophy Connected with the Subject.* Vol. 1. Reprinted from the edition of 1790 (Birmingham: Thomas Pearson). New York: Kraus Reprint Co.
Reichenbach, H. (1938). *Experience and Prediction: An Analysis of the Foundations and the Structure of Knowledge*. Chicago: University of Chicago Press.
Schlesinger, G. (1976). *Confirmation and Confirmability*. Oxford: Clarendon Press.
Taylor, J. (1982). *An Introduction to Error Analysis.* Oxford: Oxford University Press.
Will, C. (1986). *Was Einstein Right?* New York: Basic Books.
Woodward, J. (1983). Glymour on Theory Confirmation. *Philosophical Studies*, 43, 147-157.
_____. (1989). Data and Phenomena. *Synthese*, 79, 393-472.