OBSERVATIONS, THEORIES AND THE EVOLUTION OF THE HUMAN SPIRIT*

JIM BOGEN†‡

Philosophy Field Group
Pitzer College

JIM WOODWARD

Division of the Humanities and Social Sciences
California Institute of Technology

Standard philosophical discussions of theory-ladeness assume that observational evidence consists of perceptual outputs (or reports of such outputs) that are sentential or propositional in structure. Theory-ladeness is conceptualized as having to do with logical or semantical relationships between such outputs or reports and background theories held by observers. Using the recent debate between Fodor and Churchland as a point of departure, we propose an alternative picture in which much of what serves as evidence in science is not perceptual outputs or reports of such outputs and is not sentential in structure.

I

In a recent exchange,[1] Jerry Fodor (1984, 1988) and Paul Churchland (1988) revive an old debate about the theory-ladeness of observation. Our aim in this paper is to examine and to criticize a number of the assumptions that underlie this debate. We argue that the debate has failed to focus on the epistemologically relevant features of perceptual processes—which have to do with their reliability rather than their degree of informational encapsulation—and that discussion has been conducted within an excessively sentence-oriented (or propositional) framework which has impeded understanding. We conclude that this framework needs to be radically rethought.

*Received August 1991; revised November 1991.

†We would like to thank Paul Churchland, Jerry Fodor, David Hilbert and especially Cliff Hooker for helpful comments and correspondence.

‡Send reprint requests to the authors, Philosophy Field Group, Pitzer College, Claremont, CA 91711, USA.

[1]This paper focuses exclusively on the views expressed in Fodor (1984, 1988) and Churchland (1988). In a paper we only learned of after completing this essay, Fodor (1991) adopts a position regarding the role of perception in science and the significance of perceptual encapsulation which is considerably closer to (although by no means identical with) the position taken in this paper. Also, as we note below, the criticisms of sentential models of human cognition advanced elsewhere by Churchland (e.g., 1979, 1989) fit very naturally with many of the claims advanced in this paper.

Philosophy of Science, 59 (1992) pp. 590–611

II

Both Churchland and Fodor take theory-ladeness to have to do with the relationship between an observer's theoretical beliefs and what they call his "observational judgements" or "perceptual beliefs", items which they take to be sentential in structure, and capable of being true or false. For Churchland observational judgements are theory-biased because they fail to meet the following condition for theory neutrality:

> (N^c) An observational judgement is theory neutral just in case its truth is not contingent upon the truth of any general empirical assumptions, just in case it is free of potentially problematic presuppositions[2] (Churchland 1988, 180).

Churchland thinks our present level of scientific understanding is constrained by limits placed on perception by our present background theories. Fortunately, "perceptual knowledge is plastic" (ibid., 167), and by adopting better background theories we may significantly extend what we can perceive, thus achieving an important scientific advance.

Fodor agrees that perceptual beliefs are not neutral in the sense of N^c; he thinks they are affected by hardwired background theories supplemented by "background information about certain pervasive features of the relations between distal layouts and their proximal projections" (1988, 189). But such bias "leaves perception neutral with respect to almost all theoretical disputes . . . . It is therefore not the sort of bias that could . . . ground any general argument for the unreliability of observation" (ibid.). This is because perceptual judgement employs only background theory shared by most observers, and because perceptual judgements are the outputs of visual, auditory and other perceptual "modules" which are "cognitively encapsulated" (ibid., 167). This encapsulation means that the psychological units which produce perceptual judgements lack access to

[2]Churchland does not explicitly say what he means by "contingent upon" and "presuppose", but his ensuing discussion shows that he is thinking primarily, if not exclusively, in terms of logical or semantic relationships between observational judgement O, and background theory T. Thus O is "contingent upon" the truth of T if $\sim T$ entails $\sim O$. Although we will not belabor the point, we note that this characterization does not correspond to the standard notion of theory-ladeness in the philosophical literature or to the notion which Churchland tacitly employs in much of his discussion. The standard notion has to do with the way in which an observer's background beliefs, training, previous learning and so forth causally influence what he perceives or what perceptual beliefs he holds. The empirical evidence (such as inverting lense experiments, and downward pathways in the nervous system) Churchland marshals for the theory-ladeness seems to presuppose this causal notion. By contrast, on N^c, T will load O as long as it stands in the right logical relationship to T regardless of whether the observer who judges O also holds T and regardless of whether the observer's holding T causally influences his judgement that O. On N^c we know independently of empirical investigation that all perceptual judgements are laden with huge amounts of theory—for example, all my observational judgements are laden with the "theory" that not all my observational judgements are false.

theoretical beliefs other than the common background theory involved in their operation. Encapsulation insures that the outputs of the perceptual modules of different observers will not be influenced by the sorts of background beliefs over which scientists are likely to disagree. Thus, although Fodor admits that perception is theory biased according to $\mathbf{N}^c$, he thinks that most perceptual judgements are theory-neutral in the sense that they meet the following condition:

> ($\mathbf{N}^f$) Observational beliefs are theory-neutral just in case their acceptance as true depends only on the common perceptual background theory shared by almost all observers.

Churchland rejects this encapsulation thesis. Moreover, he holds that the thesis would have no epistemological significance even if it were true (1988, 181). We will examine his arguments below. What we wish to emphasize at this point, and to explore in more detail in the following sections, are the similarities between Fodor and Churchland. Both think that the issue of whether observation is theory-neutral in the sense of $\mathbf{N}^c$ or $\mathbf{N}^f$ is extremely important in understanding science and our general epistemic situation. (Churchland says his disagreement with Fodor concerns ". . . the fundamental character of the human epistemic situation, and the long-term possibilities for the evolution of the human spirit" [1988, 185].) Both think that the perceptual judgements of individual perceivers play a fundamental evidential role in science. Finally, although they disagree about how perception depends on background theory, both think theory-neutrality has to do with whether the truth of perceptual judgements (or their acceptance as true) depends upon the truth (acceptance as true) of background theory. We will argue against each of these claims.

III

Both Churchland and Fodor assume that what they call "observational beliefs" or "perceptual knowledge" play a decisive role in theory testing. What is this role? The answer would be obvious if theories were tested against reports of what individual observers perceive. But theory testing rarely has this simple structure.

In two recent papers (Bogen and Woodward 1988; Woodward 1989), we introduced a distinction between *phenomena* and *data*. It is claims about phenomena which scientific theories generally seek to predict and systematically explain and which are directly compared with theories in testing. Examples of phenomena include the melting point of lead, the occurrence of weak neutral currents, the behavior of the reticular formation of the brain during sleep and recency effects in short-term memory. Phenomena are detected and measured through the use of data (e.g., reports of individual temperature readings, bubble-chamber photographs, drawings of specimens viewed through microscopes, records of the electrical activity from single cell recordings in the central nervous system). Data are *records* or *reports*—accessible to the human perceptual system and available for public inspection. Some data (e.g., reports of measurement results written in laboratory notebooks, or the drawings of field geologists) are produced by human perceivers. But as the above examples illustrate, many data are produced by nonhuman measurement and recording devices. As we will see below, data *cannot* in general be identified with the perceptual experiences of individual observers or with reports or descriptions of such experiences or with (what Fodor and Churchland call) perceptual beliefs.

For present purposes, the point of distinguishing data from phenomena is that data constitute the observational evidence scientists use to investigate phenomena. This means that the role of perception in theory testing depends on how perception figures in the production of data and in the use of data to investigate phenomena. In the remainder of this section we suggest an analytical framework for understanding the role of perception in phenomenon detection. We begin with a discussion of what we call reliability, which we take to be the central epistemic virtue investigators desire in data and in the processes by which phenomena are detected from data.

The reliability of data has to do with its trustworthiness as a sign or indicator of the phenomenon under investigation. We think of procedures for phenomenon detection (whether experimental or nonexperimental) as consisting of two components. The first includes the causal processes which lead (or so the investigator hopes) from the phenomenon of interest to the production of data—for example, from the occurrence of certain particle interactions to characteristic bubble-chamber photographs. The investigator tries to arrange matters so that, depending upon which of several claims concerning the phenomenon is true, distinctively different data will be produced. The reliability of this component of the detection process depends on, for example, the physical characteristics of the general experimental design or layout—including the possible presence of background noise or confounding causes of various kinds—the adequacy of steps taken to eliminate or control for such factors, and the physical characteristics of the instrument or detection device employed. We call this component *data production*.

In the second component of detection, an investigator (or, more typically, a group of investigators) moves from data to conclusions about phenomena—for example, from bubble-chamber photographs to the conclusion that neutral currents occur. We call this component *data interpretation*. This component includes techniques of data analysis and reduction, including statistical procedures and decisions to discard data or smooth or correct it in various ways. It also may include the explicit use of background or theoretical assumptions, as when kinematic assumptions are used to calculate the momentum of some particle from a bubble-chamber photograph. It may also include arguments by which the investigators try to determine whether various sources of error have been adequately controlled for, or try to correct for known and suspected sources of error by calculation. We emphasize the extraordinary variety of considerations that may be involved in this interpretive process. As noted, some may involve explicit formal calculations and derivations. In other cases, for example, in deciding whether to discard data or whether certain sources of error have been adequately controlled for, the relevant considerations may be partly informal and nonalgorithmic—matters of "judgement" or experimental skill or diligence, rather than explicit calculation in accord with formal rules (see Galison 1987).

The reliability of data or any particular technique for phenomenon detection will turn on whether the various elements in the detection process work together to allow investigators to correctly discriminate[3] among competing claims regarding the phenomenon of interest. Typically investigators will have in mind a specific set of such alternative claims and will try to show that their data and detection procedures preclude (or at least render implausible) every alternative except the claim they wish to establish. Thus, the neutral current experiments and the bubble-chamber photographs they produced will be reliable if they can be used to distinguish cases in which neutral currents occur in the detection apparatus from various alternative possibilities in which no such interactions occur—for example, from the alternative possibility that the data are caused by high energy neutrons (which in a bubble chamber can mimic many of the effects of neutral currents; see Galison 1983). The reliability of data will thus be relative to the processes and techniques used to produce and interpret the data and the possibilities to be discriminated. The same data may be reliable relative to one interpretive procedure or set of alternative claims but unreliable relative to another interpretive procedure or set of alternative claims.

While reliability is a matter of correct discrimination, we do not believe that there is some general readily specifiable set of necessary and sufficient conditions which data must satisfy if they are to be used to discriminate correctly. In particular, we deny that the use of data to correctly

[3]"Correctly discriminate" is shorthand for "discriminate in a way that satisfies appropriate epistemic goals and values". We assume that these may include, but are not necessarily limited to, discriminating truths from falsehoods—thus one might also aim at discriminating among competing claims with respect to their usefulness for purposes of prediction and control.

discriminate between competing phenomena claims requires that data stand in some distinctive logical relationship to those claims of a sort that has been the subject of standard philosophical treatments of confirmation. For example, we would reject naive versions of the H-D model according to which data E discriminate in favor of alternative H_1 and against alternative H_2 only if H_1 entails E and H_2 entails $\sim E$. We also reject the suggestion that E discriminates in favor of H_1 and against H_2 only if $P(H_1/E) >> P(H_2/E)$. Although these suggestions (or more sophisticated variants) may successfully model some of the arguments to which investigators appeal when they use data to discriminate among competing claims, they do not state generally necessary conditions for correct discrimination, and often the conditions they state are not sufficient either. On our view, the mechanisms, processes and patterns of argument that underlie correct discrimination are highly disparate and heterogeneous, and many have little to do with the logical and probabilistic relationships which have figured so prominently in philosophical discussions of evidential support. As an alternative paradigm, consider perceivers who achieve a low long-run error rate in an object recognition task like distinguishing crows from ravens. It is doubtful that perceivers do this by recognizing that some sententially expressible features E stand in some appropriate confirmation-theoretical relationship to hypotheses about bird type. This example shows that correct discrimination does not always require instantiation of some familiar logical or probabilistic relationship between evidence and hypothesis. Formal models of confirmation may tell us equally little about some forms of reasoning from data to phenomena.[4]

We will not try to provide a single characterization of what reliability consists in, but one aspect of reliability is both general and tolerably clear. Suppose that we can provide a nontrivial characterization of what it is to repeat some detection process and that the process possesses stable, empirically ascertainable error characteristics under repetition. Then the reliability of the process can be characterized in terms of its long-run error characteristics. A reliable process will be a process having a high probability of outputting correct discriminations among a set of competing phenomenon claims and a low probability of outputting incorrect discriminations. By extension, reliable data will be data which are produced by or contribute to a reliable detection process. Alvin Goldman's (1986) reliabalism is a familiar and fruitful application of this notion of reliability to traditional epistemology and a similar idea plays an important role in many areas of scientific investigation. For example,

[4]In Bogen and Woodward (forthcoming) we criticize in more detail what we call the IRS conception of evidential support—the view that the relation between hypothesis and evidence not always be understood in terms of Inferential Relations between Sentences.

as already intimated, often one can investigate empirically whether human perceivers can make certain perceptual discriminations reliably where this has to do with the frequency of correct discriminations under repeated trials. Determinations of personal error rates in observational sciences like astronomy make use of this understanding of reliability (see Boring 1950, 134–153.) Similarly an automated data reduction procedure which sorts through batches of photographs selecting those satisfying some preselected criterion is operating reliably if it actually classifies the photographs according to the indicated criterion with a low error rate.[5]

For reasons we lack space to explore, we doubt that all considerations bearing on the reliability of particular uses (or tokens of) detection processes can be captured within this framework of the error characteristics of repeatable processes.[6] Nonetheless, we think that the long-run error characteristics framework captures one central aspect of reliability and we will exploit this below. In particular, we will see that the notion of reliability as low long-run error rate allows us to make sense of the possibility that data can be reliable even though they (or their most natural representations) lack truth-conditions and do not stand in any logical or semantical relationships to background theory.

What role does perception play in this general framework? There are two different points at which human perceivers can figure in phenomenon detection. Sometimes the perceiver herself will be part of the causal chain or process leading to the *production* of data. Here what the perceiver outputs will be either the data or an intermediate step in its production. For example, when a geologist looks at a rock formation and draws a diagram of it or when Galileo looked through his telescope and drew pictures of the moon, their activities were parts of causal chains which produced data (diagrams, pictures). Note, however, that in many areas of science, data will not be produced in this way, but rather by the operation of nonhuman instruments.

Second, whether or not human perceivers figure in data production, the *interpretation* of data (its use to support claims about phenomena) will virtually always involve human perception. Thus, in experiments designed to detect neutral currents, the data were bubble-chamber photo-

[5]The claim that a detection process is reliable is always an empirical claim whose truth will often depend upon local details of the detection processes. In addition to information about long-run error characteristics, grounds for belief in reliability typically involve considerations of replicability, consistency with other detection techniques, calibration techniques, and behavior in response to manipulation. We also emphasize that assessments of reliability need not require possession of a detailed explanation of how the detection processes works. See Bogen and Woodward (1988) and Woodward (1989).

[6]For further discussion, see Bogen and Woodward (forthcoming), which argues that assessments of reliability often depend upon a kind of causal analysis aimed at eliminating or correcting for confounding influences in data production. Information about long-run error rates is not the only kind of information relevant to such assessments.

graphs that were not the output of human perceptual processes. Nonetheless, human perceivers had to look at such photographs, discriminate among them, draw certain photographs to the attention of others, recognize characteristic patterns, and measure tracks as part of the complex series of procedures and arguments which led to the detection of neutral currents (Galison 1985).

In thinking about the role of perception in science and the influence of theory on perception, these two ways in which perception may figure in phenomenon detection must be kept distinct. Philosophical treatments of theory-ladeness like Fodor's and Churchland's do not apply to data-production in which human perception plays no role. The characteristics of the causal processes leading to the production of bubble-chamber photographs in some experiment are relevant to the experiment's reliability but neither these processes nor the photographs are "theory-laden" in the sense in which Fodor, Churchland and most other philosophers use that expression. Issues about the theory-ladeness of perception are relevant to such cases only insofar as they are relevant to the role of perception in data-interpretation. While we lack the space for a detailed discussion, our view is that although perception (and worries about the reliability of perception, and the control of perceptual error) are sometimes central to data-interpretation, the reliability of the procedures used to move from data to phenomena typically has more to do with nonperceptual (e.g., statistical) techniques of data analysis, and with strategies for the control of nonperceptual error. The procedures involved in data-interpretation are simply too heterogeneous and, in many cases, too different from those involved in perception for it to be analytically illuminating to assimilate them all to the latter category. (For further defense of this claim, see Bogen and Woodward 1988.)

Whether perception figures in phenomena detection at the stage of data-production or data-interpretation, its contribution to the reliability of the overall detection process can be empirically investigated. Here are two illustrations.

First, an important experimental controversy between Rutherford and Chadwick in Cambridge and Petterson in Vienna turned on the ability of trained observers to perceptually distinguish and count flashes of varying brightness on a scintillation screen (Steuwer 1985). The data were the counts the observers produced. Petterson attempted to use these to discriminate between protons and alpha particles and thus to determine whether carbon atoms emit protons under bombardment by alpha particles (the phenomenon of interest). Experimental investigations of the conditions under which human observers could reliably distinguish scintillation flashes (that is, empirical determination of error rates) made the Cambridge group believe this method of detection was unreliable. During a visit to Petterson's laboratory, Chadwick confirmed this; Petterson's counters claimed to observe scintillations even when (unknown to them) Chadwick removed the particle source or introduced absorbing material exceeding its calculated range.

Second, in many areas of science, the visual recognition of complex patterns plays an important evidential role. For example, data consisting of written reports, drawings, or diagrams produced as a result of visual recognition of rock types or geological structures can serve as important evidence for geological hypotheses. In such cases we can think of reliability as having to do with long-run error characteristics—with the error rates with which investigators will correctly discriminate and classify different objects and structures. As historians like Rudwick (1985) emphasize, reliability in such recognition tasks typically requires considerable training and exposure with feedback to the structures to be classified. Reliability in such activities thus may vary considerably across individuals (including investigators in the same field) and may be highly task-specific: The same person may be reliable with respect to one sort of discrimination or recognition task, but not with respect to another. As Rudwick shows, the scientific value of the data a particular investigator produces will turn in part on rough and ready empirical assessments of his reliability characteristics by his peers. That is, the evidential significance accorded to the data an observer produces will vary with his reputation for reliability in the long-run error characteristics sense described above.

We introduce these examples to convey a sense for how perception and issues about the reliability of perception actually figure in real-life scientific investigations. One of our central themes is that the fundamental question epistemologists should ask about the role that perception plays in some detection process is not whether perception is theory-neutral or theory-biased, or plastic or informationally encapsulated, but rather, whether it is reliable, and whether it contributes or detracts from the overall reliability of the detection process in which it figures. As we will see, much of what affects the reliability of perception is not captured by $\mathbf{N}^c$ or $\mathbf{N}^f$ or the arguments which divide Fodor and Churchland.

We said above that although perception can play a causal role in data-production and interpretation, data are not reports or descriptions of the subjective experiences of individual perceivers or reports of individual perceivers' beliefs or judgements about how matters seem perceptually to them. Part of our grounds for saying this should already be apparent. Many data are not produced by human perception. Even when data are produced by perception, they often consist of pictures or diagrams which do not have the sentential or propositional structure philosophers associate with belief and judgement.[7] Moreover, even when data are produced by human perceivers and have a sentential or propositional structure, the view that data describe subjective experiences of individual perceivers is still mistaken. We stress this because a great deal of philosophical discussion of theory-ladeness (including much of Fodor's and Churchland's discussion) presupposes that true perceptual beliefs representing how matters seem to individual perceivers play a central evidential role in theory testing and hence that it is important to ascertain how background theory may affect such beliefs. But pervasive as this idea is among philosophers, we will see that it is by no means obvious how to fit it into a descriptively accurate picture of scientific practice. It is data rather than perceptual beliefs that play a central evidential role in science and data are typically not descriptions of perceptual appearances or reports of perceptual belief at all.

Consider an observer who makes measurements by looking at a dial and reporting the result. His visual experiences should influence the data he writes down. But the epistemic value of this data will depend upon their contribution to the reliability of the detection and measurement processes they figure in, and not on whether they accurately describe the observer's perceptual experiences. Indeed, data which provide autobiographically correct or phenomenologically detailed records of the observer's subjective perceptual experiences can detract from the reliability of the detection processes in which they figure. An observer who knows his perception is unreliable, for example, because of limitations in sensory acuity or reaction time, may produce epistemically better data by correcting for these factors in his reports instead of describing his visual experiences without correction. Cases of this sort are common in observational astronomy. Thus data recording the time when a star passes a certain position may not be intended to represent the time at which it appears to the observer to pass, if the observer knows she makes some systematic error. Outside of perceptual psychology, the production of accurate portrayals of how things look to individual observers is seldom the object of data production (see Gregory 1981 and Sheehan 1988). Instead investigators want data which allow them to discriminate reliably among different phenomena claims.[8]

[7]The central role played by nonpropositional forms of representation in science is increasingly recognized by historians and sociologists, but so far has received little attention from philosophers. See, for example, the papers in Lynch and Woolgar (1990). The implications of this for standard models of evidential support which are framed in terms of logical or probabilistic relationships between sentential structures deserve further exploration.

[8]Paul Feyerabend (1969) argues for the "possibility" that one might test a hypothesis without undergoing "some relevant [sensory] experience" (791–792). Feyerabend's emphasis on the irrelevance of the subjective or phenomenal elements of experience is similar

Another reason not to suppose data must describe how matters appear to individual perceivers derives from the public character of data and the social character of the scientific activities in which data figure. Scientists appeal to data to persuade other scientists of the correctness of their claims about phenomena. For data to function in this way, it is crucial that they be produced in a form in which they can be publicly inspected, checked by replication of relevant experiments and observations, and readily classified, aggregated, and analyzed by other investigators. As a number of recent historical studies (e.g., Swijtink 1987, and Shapin and Shaffer 1985) have emphasized, learning how to produce data having these features, and thus not too closely reflecting idiosyncratic features of the perceptual experiences of individual observers, is often a difficult and laborious undertaking, and success in this enterprise an important scientific achievement. The preference, within many areas of science, for using instruments in place of human observers wherever possible and, where not possible, for minimizing, systematizing and regularizing the role of the observer, reflects the fact that data produced by instruments is more often publicly assessable than data produced by human observers. Even when data are produced by human perceivers, they may be deliberately produced in a form which fails to portray exactly the perceiver's perceptual experiences (or even the object which is perceived) in order to make them more readily and publicly checkable. For example, drawings in geology and neurobiology may exaggerate certain dimensions and compress others, simplify boundaries and omit contours and a great deal of visually prominent detail in order to facilitate use by others.

IV

In the remainder of this paper, we apply our analytical framework to Churchland's and Fodor's treatments of theory-bias. This section focuses on the common sentential orientation shared by $\mathbf{N}^c$ and $\mathbf{N}^f$. Both Churchland and Fodor treat perceptual outputs as items with a sentential structure and both think that the truth of such items and their logical or semantical relationships to background theory are crucial to their evidential status. Against these common assumptions, we will argue that

1. The outputs of human perception that are scientifically relevant often lack sentential or propositional structure and therefore cannot have truth-values.

to ours, but does not go far enough. What Feyerabend describes is not just a "[logical] possibility", but is rather an accurate description of much present-day science. More generally, we depart from Feyerabend in emphasizing the public, social and often nonpropositional character of data. By contrast, it is crucial to Feyerabend's conception of theory-ladeness that both evidence and theory be sententially representable.

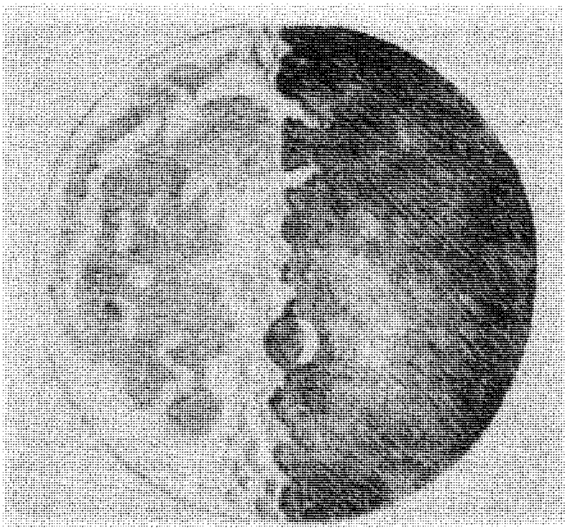

Figure. Moon. From *Sidereus Nuncius* by G. Galilei; translated by A. Van Helden. Copyright © 1989 by the University of Chicago Press. Reprinted by permission.

2. Even when perceptual outputs are such that the notions of truth and falsity apply, they usually need not be literally true to play the epistemological role required of them in scientific reasoning. In addition, nonpropositional perceptual outputs need not possess some truth surrogate like exact representational accuracy. Moreover, the truth of an investigator's perceptual beliefs is frequently irrelevant to the evidential value of the data which figures in his investigations. It is therefore epistemologically irrelevant to worry about how the truth of such items depends upon background theory, as $\mathbf{N}^c$ and $\mathbf{N}^f$ suggest.
3. The explanation for (1) and (2) is that the items which figure as input, output or intermediate stages in a detection process need not be propositional in structure or possess a feature like truth or representational accuracy for the detection process to be reliable. We also suggest in the following section that the "theory" which figures in a detection process need not be true, or even the sort of thing which could be true or false, for the detection process to be reliable.

Consider the pictures Galileo drew on the basis of observations with his telescope and used to argue that the moon is not a smooth, crystalline sphere (see figure).

These drawings lack sentential structures and truth-conditions. The well-known seventeenth-century controversies over the drawings did not concern their truth. Instead they had to do, for example, with whether various features represented in the drawings were artifacts caused by imperfections in Galileo's telescope. These are issues about the *reliability* of Galileo's data (and the processes which produced them) not about the truth of his data. Beliefs about the reliability of Galileo's drawings and the factors which may have influenced their reliability certainly have truth-values. But these beliefs about reliability are not perceptual beliefs or outputs in the sense intended by Fodor and Churchland. Such beliefs about reliability should not be confused with Galileo's data themselves, which were pictures.

What if anything is the relevance of Galileo's perceptual beliefs to the reliability of his drawings? According to our previous discussion, Galileo's drawings were reliable if they could be used to correctly discriminate between the two hypotheses that interested him—that the moon is a smooth, crystalline sphere and its denial. Because the exact number of craters or mountains on the moon, their sizes, or the distances between them, and so on, are irrelevant to this discrimination, the truth-values of Galileo's perceptual beliefs (if any) about such matters are also irrelevant to the reliability and evidential value of the drawings. Thus even if false perceptual beliefs—for example, about the relative distances between mountains or craters—figured in the causal process that produced Galileo's drawings, this would not detract from their reliability. Indeed, it is not even clear that Galileo needed to hold the perceptual belief that (**R**) the surface of the moon looks rough and irregular in order for his data to be reliable. Of course Galileo probably believed **R**, but as a matter of historical fact neither this belief nor any report of Galileo's holding it were regarded as evidence that the moon's surface was irregular and noncrystalline. It was not Galileo's perceptual beliefs or reports about his beliefs but rather his *drawings* which served as data, for only they had the publicly assessable and checkable character required for the evidential role of data. Moreover, the fact that reliable data can be produced by nonhuman instruments also shows that true perceptual beliefs are not required for reliability.

Examples about which similar points could be made can be found in many areas of science in which pictures, drawings and graphs are common. Since such data lack truth-values, we cannot even ask what their truth or acceptance as true depends upon, as $\mathbf{N}^c$ and $\mathbf{N}^f$ require. Although such outputs can be reliable or unreliable, we cannot capture what their reliability consists in or what evidential role they play by invoking the difference between true and false perceptual beliefs and the way in which their truth (acceptance) depends upon background theory.

The following example, involving neither pictures nor diagrams, provides another reason why focusing on truth-values (of data, of sentential representations of data, of perceptual beliefs associated with data) may not tell us much about the evidential value of data. Here the unimportance of truth is a consequence of the fact that data that are full of noise, error and other sources of falsity and inaccuracy can nevertheless be used to reliably discriminate among conflicting phenomena claims.

Priestley devised a test (also used by Lavoisier) to show that an unknown gas emitted from heated oxides of mercury, iron, and other metals differs from atmospheric air. Anachronistically described, the test depends on the fact that oxygen combines with nitric oxide to produce water soluble nitrous oxide. Measured amounts of nitric oxide and the gas to be tested are combined over water in an inverted, graduated tube. As the nitrous oxide thus produced dissolves, the gas volume decreases and the water rises in the tube. At fixed volumes, the more uncompounded oxygen a gas contains, the greater its decrease in volume, as measured by watching how far the water rises. In their first experiments the volume decreases reported by Priestley and Lavoisier did not discriminate between the hypothesis that the unknown gas was oxygen and the hypothesis that it was atmospheric air. The data from later experiments report decreases that clearly distinguish the two hypotheses, but the evidential superiority of these data is not a result of their being exactly true or accurate. In their earlier experiments Priestley and Lavoisier did not use enough nitric oxide to combine with all the oxygen released from the heated oxide. Because oxygen does not dissolve readily in water, the volume of gas in the graduated tube did not drop nearly as much as it would have if enough nitric oxide had been used to take all of the oxygen up into the highly soluble nitrous oxide. In the later experiments Priestley and Lavoisier used enough nitric oxide to produce an unmistakably greater volume change than could have been obtained with atmospheric air. This difference in reliability is obviously not a matter of the later data exactly coinciding with a true volume decrease to which the former data fail to correspond. The nature of the equipment and techniques available for measuring the gases, introducing them into a graduated tube, and observing volume decreases virtually guaranteed considerable error in the second (reliable) body of data as well as the first (unreliable) body (see Conant and Nash 1957, 75; Holmes 1985, chap. 2; and Priestley 1970, 23–41).

The framework provided by Churchland and Fodor does not capture what is epistemically relevant here. Suppose, for the sake of argument, that the following counts as a perceptual belief:

> (J) The observed result of adding 3 volumes of "nitric oxide" to 2 volumes of the unknown gas was to reduce the latter to 1 volume (see Conant and Nash 1957, 101–102).

We have seen that for this experiment to be reliable **J** need not coincide with, or accurately represent, the true volume decrease. Moreover, **J** need not be literally true even when understood as a report of the measurement result that the experimenters perceived and Priestley and Lavoisier need not have accepted **J** as true. Suppose the actual measurement result was closer to 0.9 than to 1. Suppose Priestley and Lavoisier suspected this, disbelieved **J**, but considered further precision unnecessary. None of these inaccuracies or failures of literal truth in **J** matter as long as their data can be used to reliably make the discrimination that interests them.

Someone might suggest that even if **J** need not be exactly true, reliability requires it to be "approximately true", and that, accordingly, theory-ladeness could be characterized as dependence of the approximate truth of the perceptual belief (or its acceptance as approximately true) on background theory. This suggestion is unilluminating because the required notion of approximate truth cannot be given a general, context-free characterization. Even when data have a natural sentential representation, the degree of nearness to truth required in data is relative to what the data are to be used for and how they are to be used. That is, the required feature can only be characterized by reference to the competing phenomena claims under consideration and the methods by which the data will be used to discriminate among them. In some contexts data which are not in any intuitive sense approximately true permit perfectly reliable discrimination. In other contexts data which are in an intuitive sense much closer to truth may fail to discriminate reliably. Any notion of approximate truth that captures the notion of evidential value of data will, we suspect, reduce to our notion of reliability.

A similar point holds for nonsentential data like drawings and diagrams. We have seen that these need not be exactly representationally accurate to serve as data. Here too what counts as an acceptable approximation to representational accuracy depends on the possibilities to be discriminated and the methods by which the data will be used to discriminate among them. To call a diagram approximately true is just a misleading way of saying that it can be used for reliable discrimination. Such talk offers no independent explanation of what reliability is.

The application of all of this to Fodor and Churchland should be obvious. Because reliability does not require truth (approximate truth, exact representational accuracy, and so on) of data or perceptual beliefs, the evidential value of such items cannot turn on how their truth (representational accuracy) depends upon background theory, as $\mathbf{N}^c$ and $\mathbf{N}^f$ suggest. By contrast, our framework provides a straightforward treatment of the examples just considered, and many of our heterodox sounding claims about data follow naturally from it. For data to be reliable, they need only make an appropriate causal contribution to a process enabling investigators to discriminate correctly among competing phenomena claims. To play this causal role, data need not be perfectly accurate, or error free or even sententially representable. Fodor's and Churchland's frameworks require that background beliefs stand in logical or semantic relationships to the perceptual beliefs they suppose serve as evidence. This approach gets into difficulty because, among other things, data often lack the syntactic and semantic properties needed for these relationships to hold. Our approach focuses instead on the causal influence of various factors on the reliability of processes of data-production and interpretation, and thus does not require logical or semantic relationships between data and background theory.

V

In this section we consider Fodor's encapsulation thesis and argue that it is largely irrelevant to understanding the role of perception in science.

The outputs of Fodor's perceptual modules have to do with how things perceptually appear to an individual perceiver—for example, that the lines in the Mueller-Lyre illusion look unequal in length. Fodor emphasizes that this appearance is not penetrated by other kinds of background information—the lines still look unequal after we learn they are equal. He thinks such facts about nonpenetration help explain how observers with different background beliefs can reach a consensus about what they observe which can serve as an evidential basis for science. However, there is a fundamental problem with Fodor's account: *The outputs of the perceptual modules (how things look) or reports about such outputs are usually not what serves as evidence in science.* What serves as evidence are data. What scientists need to reach consensus about is not the perceptual judgements of individual observers, but the reliability and the evidential significance of data.

To illustrate, consider a measurement procedure in which an observer must ascertain whether a pointer on a balance coincides with a certain marking on a scale behind it. Suppose the observer views the pointer from a position at a sharp acute angle to the plane of the pointer, rather than looking directly down from a position normal to this plane. Suppose also, as seems likely, that informational encapsulation holds—viewed from this position, the pointer appears to coincide with a certain marking on the scale, regardless of the observer's background beliefs, including those about whether the apparent position of the pointer coincides with its true position. Suppose different observers reach consensus about this appearance—the position of the pointer looks the same to all when viewed from this acute angle, regardless of their background beliefs. To a first approximation, all of this is epistemically irrelevant; what matters epistemically is the reliability of this procedure and the data it produces, not how things look phenomenologically to the investigators. In fact, such sighting along an acute angle yields unreliable measurements and one must look directly down from a position normal to the plane of the pointer to reliably ascertain the pointer's position visually. A common device for insuring this is a small mirror placed under the pointer. When the reflection of the pointer coincides with the pointer itself, one is in the correct viewing position.

This example illustrates how what is epistemically crucial is the reliability of perception, rather than its informational encapsulation or lack thereof. Even if all perceptual judgements were as informationally encapsulated as Fodor claims, judgements about reliability (and the kinds of considerations which ground such judgements) are plainly not. Indeed, the epistemic credentials of judgements about reliability are usually *improved* by extensive use of background information, by cross-checking and calibrating techniques designed to detect the same result, and by communication and comparison involving different investigators. The achievement of a scientifically relevant consensus requires, not agreement in perceptual judgement in Fodor's sense, but publicly assessable data and agreement in judgements of reliability. Even if Fodor's claims about perceptual encapsulation were correct, they would explain little about the achievement of consensus and reliability in science since they neglect the evidential role played by data and by judgements of reliability.

The following case in which differences in individual perceivers actually did produce discrepancies in perceptual data underscores the epistemic irrelevance of Fodorian encapsulation and further establishes that agreement in perceptual judgements about how things look is not essential to the establishment of scientific consensus and reliability. In a famous investigation of discrepancies in stellar transit times reported by different observers, F.W. Bessel found that despite some variability in each individual's observations, systematic and relatively stable differences existed between observers (Gregory 1981, Boring 1950). This made it possible to devise a personal equation for each individual, relating his observations to those of other observers. The personal equation could be used to "calibrate individual observers with respect to one another . . . [and] compensate for individual differences" (Gregory 1981, 212). Every observer could be "corrected" by averages taken from other observers (ibid., 213). Originally, when no method other than human observation could be used to determine absolute transit times, personal equations could give only relative differences between observers, and could not correct for absolute errors. The subsequent development of electromagnetic detection and recording devices which do not rely on human observation to produce data made possible the determination of "absolute" errors for each observer.

The use of personal equations makes possible the comparison and correction of different perceptual reports in a way that considerably enhances their reliability as evidence. Such equations illustrate how even in the presence of different perceptual judgements by different observers (judgements which are probably at least in part hardwired or informationally encapsulated), appropriate cross-checking and calibration can nonetheless make possible achievement of consensus and reliability. They thus show again how little encapsulation and agreement about how things look matter epistemically.

Next consider cases in which a group of experts produce data (e.g., geological drawings) which depend on a learned visual pattern recognition skill. The perceptual mechanisms involved here are clearly heavily influenced by prior training but this makes no difference to the achievement of consensus or reliable knowledge, as long as the data produced are publicly assessable and the experts are reliable. Even untrained or inexperienced investigators who cannot see what the experts see or make the discriminations the experts make may still justifiably use the experts' identifications of rock types and geological structures. Once again, we see that agreement about how things look is not what matters epistemically. Moreover, like the last two examples, this shows there is simply no systematic relationship between the extent to which perception, as used in a particular detection task, is reliable and the extent to which it is informationally encapsulated. Sometimes extensive top-down processing enhances reliability and sometimes it diminishes it. Similar conclusions hold for highly modular mechanisms. Thus the general claims about plasticity and encapsulation dividing Fodor and Churchland do not isolate what is epistemically relevant about perception—its bearing on reliability in specific situations.

We turn now to Churchland's views about perceptual encapsulation. Churchland has two objections to encapsulation. First, Churchland claims against Fodor that changes in background belief often change what Churchland calls the character of perception itself—that is, how matters perceptually seem. Second, Churchland claims that even if true, the encapsulation thesis is epistemologically irrelevant because theories are compared not with (what Churchland calls) "sensations", but rather with sentences or propositions expressing observational judgements or beliefs. The meanings of such observation sentences depend, Churchland thinks, upon the background beliefs or "conceptual system" in which they are embedded and to which they stand in various semantical and logical relationships. Because our conceptual system is plastic, the semantics of observation sentences must be equally plastic and open to penetration by background information. Thus if theories are tested against observation sentences the meaning of these sentences will be affected by background theory, regardless of whether the encapsulation thesis is true. The trouble with this argument is that because data are often nonsentential they often lack the semantical properties (e.g., truth, falsity, semantic meaning) Churchland's objection requires. Quite apart from this, we have already seen that such properties (and a fortiori how they may be affected by shifts in background theory) are often irrelevant to the evidential status of data.

Next, consider Churchland's general emphasis on how improvements in background theory can improve the epistemic credentials of perceptual belief. Applied to the pointer example, Churchland's idea should be that one makes a scientific advance by changing one's background theory in such a way as to replace the original false perceptual judgement about the position of the pointer by a true perceptual judgement. It is striking that this is *not* how reliability is actually achieved in this example. Reliability is improved by changing the *causal* conditions under which perception occurs (by changing the position of the perceiver), the causally relevant characteristics of the instrument (e.g., introducing the mirror), and in this way, changing the reliability characteristics of the detection process in which perception figures. More generally, reliability is enhanced by altering the character and hence the reliability of the data-*production* process which is in part external to the observer, not (just) by improving the observer's background beliefs or interpretive scheme. Moreover, as our discussion of Bessel illustrates, many advances in reliability come, not by improving perception at all (and still less by loading it with a better theory), but rather by replacing perception entirely with mechanical detection and recording devices, or by redesigning the detection process so that perception plays a less central role. Churchland repeats Fodor's mistake of focusing too much on perceptual judgements of individual perceivers and on how changes in background belief may affect these and thus overlooks other, more effective strategies for improving reliability. While Churchland is right to say that encapsulation is epistemologically irrelevant, his account of why concedes too much to Fodor's framework and to traditional empiricism.

We turn next to a point that complements our criticisms of the sententialist orientation of $\mathbf{N}^c$ and $\mathbf{N}^f$ in section 3. Suppose extensive training and experience with feedback about successes and errors improves a medical diagnostician's ability to reliably visually identify certain patterns in X-ray photographs. It is doubtful this improvement is a matter of replacing a background theory which committed the perceiver to false perceptual judgements with a true (or more nearly true) theory permitting him to make true perceptual judgements. Connectionist or neural network models provide an alternative account of what is involved in such visual pattern recognition which arguably assigns no causal role to sentential representations. In such models, pattern learning is not a matter of changing one's background theory in the sense of sentences held to be true. Since Churchland has argued in an illuminating way for the relevance of such models to epistemology and philosophy of science in other contexts, (e.g., Churchland 1989) it is puzzling that his discussion of theory-ladeness (Churchland 1988) is conducted so exclusively within a sententialist framework. *If connectionist models provide good accounts of aspects of perceptual recognition and identification which are important in science and if they have the antisententialist implications for which Churchland has argued elsewhere, then it is a mistake to try to capture the idea that learning and training, (and similarly, one would suppose, factors like expectations and mental set) can influence what is perceived or what is perceptually recognized in terms of the sentential characterizations provided by* N^c *and* N^f. This is just the conclusion we reached on independent grounds in section 3 by focusing on perceptual outputs.

Unlike $\mathbf{N}^c$ and $\mathbf{N}^f$, our framework provides a natural way of thinking about the epistemological role of factors like training and experience in perception. In effect we take the epistemically relevant features of such factors to depend upon their causal impact on the error characteristics and reliability of perception and other detection processes. This allows us to recognize that many different factors (e.g., training, mental set, sensory acuity) that are not true or false can nonetheless be epistemologically important in virtue of their causal influence on reliability. Indeed, our view is that even when a perceiver holds a background theory of a sort that could be true or false, all that matters epistemically is how his holding the theory (causally) affects reliability; the truth of the theory or its logical or semantical relationship to perceptual output makes no difference apart from this. This explains why, as Laymon (e.g., 1983) and Hooker (1987) demonstrate, false theories can figure in reliable detection processes, and a detection process can be reliable even if laden by a theory which is not logically independent of the alternatives it is used to discriminate among.

Finally, let us briefly consider the overall motivation for Churchland's claims about perceptual plasticity in light of the above remarks. Churchland wants to establish that (1) what we can learn about nature is not constrained by what we presently perceive and that undreamt of epistemic possibilities are open to us. He argues for (1) by insisting on (2) the plasticity of perception and the possibility of changing what we perceive by appropriately changing our background theory. Thus he recommends replacing the present old-fashioned framework in which, for example, we "observe [i.e., perceive] the western sky redden" (Churchland 1979, 30) with a more scientifically up-to-date framework in which we "observe [i.e., perceive] the wavelength distribution of incoming solar radiation shift towards the large wavelengths" (ibid.). We also endorse (1), but we think its truth does not require (2), and that the expansionist treatment of perception implicit in (2) is implausible and unmotivated. As we have emphasized, the epistemic significance of perception has to do with its reliability, not with its distinctively phenomenal or subjective experiential character. If this is correct, nonperceptual techniques of measurement and detection are just as epistemically valuable as perceptual techniques as long as they are reliable. The distinction between what can and cannot be perceived thus corresponds to nothing of fundamental epistemological interest. As long as a phenomenon can be reliably detected or measured, it does not matter whether it can be perceived. Churchland's discussion assumes there is something epistemically special about perception—for example, that something of epistemic importance would be lost if we would not *see* the wavelengths of light at dusk (whatever this might mean) but could only reliably measure them. He thus ties success on our epistemic journey very closely to the plasticity of our perceptual faculties. Our contrary view is that even if perception turns out to be far less plastic than Churchland hopes, the possibilities for learning about nature would be undiminished as long as reliable measurement and detection are possible. In short, the belief that weighty epistemological matters turn on whether perception is plastic rests on a mistaken idea about the importance of perception, as opposed to nonperceptual detection, in science.

We urge in conclusion that philosophers who are interested in theory-ladeness shift their attentions away from the issues about the logical and semantical relationship between theory and perceptual judgement and about perceptual plasticity that have dominated recent discussion. If one wants to understand the role of perception in science and in the "human epistemic situation" (Churchland 1988, 185) there is simply no substitute for detailed, case-by-case empirical investigations into the reliability of perception as used in particular processes of phenomenon detection, the particular factors that affect reliability, and the strategies that are available for checking and improving reliability. This is what scientists do when they have serious concerns about the epistemic credentials of perception in scientific contexts (think of Chadwick and Bessel) and naturalistically minded philosophers of science would do well to emulate them.

REFERENCES

Bogen, J. and Woodward, J. (1988), "Saving the Phenomena", *Philosophical Review 97:* 303–352.

———. (forthcoming), "Evading the IRS", in N. Cartwright and M. Jones (eds.), *Varieties of Idealization*. Holland: Rodopi Publishers.

Boring, E. (1950), *A History of Experimental Psychology*. 2d ed. New York: Appleton-Century-Crofts.
Churchland, P. (1979), *Scientific Realism and the Plasticity of Mind*. Cambridge, England: Cambridge University Press.
———. (1988), "Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor", *Philosophy of Science 55:* 167–187.
———. (1989), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.
Conant, J. B. and Nash, L. K. (eds.) (1957), *Harvard Case Histories in Experimental Science*, vol. 1. Cambridge, MA: Harvard University Press, pp. 65–116.
Feyerabend, P. (1969), "Science Without Experience", *Journal of Philosophy 56:* 791–794.
Fodor, J. (1984), "Observation Reconsidered", *Philosophy of Science 51:* 167–187.
———. (1988), "A Reply to Churchland's Perceptual Plasticity and Theoretical Neutrality", *Philosophy of Science 55:* 188–198.
———. (1991), "The Dogma that Didn't Bark", *Mind 100:* 201–220.
Galilei, G. ([1610] 1989), *The Sidereal Messenger*. Reprint. Translated by A. Van Helden. (Originally published as *Sidereus Nuncius*. Venice.) Chicago: University of Chicago Press.
Galison, P. (1983), "How the First Neutral Current Experiments Ended", *Reviews of Modern Physics 55:* 477–509.
———. (1985), "Bubble Chambers and the Experimental Workplace", in P. Achinstein and O. Hannaway (eds.), *Observation, Experiment and Hypothesis in Modern Physical Science*. Cambridge, MA: MIT Press, pp. 309–373.
———. (1987), *How Experiments End*. Chicago: University of Chicago Press.
Goldman, A. (1986), *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.
Gregory, R. (1981), *Mind in Science: A History of Explanations in Psychology and Physics*. Cambridge, England: Cambridge University Press.
Holmes, F. L. (1985), *Lavoisier and the Chemistry of Life: An Exploration of Scientific Creativity*. Madison: University of Wisconsin Press.
Hooker, C. (1987), *A Realistic Theory of Science*. Albany: State University of New York Press.
Laymon, R. (1983), "Newton's Demonstration of Universal Gravitation and Philosophical Theories of Confirmation", in J. Earman (ed.), *Minnesota Studies in the Philosophy of Science*. Vol. 10, *Testing Scientific Theories*. Minneapolis: University of Minnesota Press, pp. 179–199.
Lynch, M. and Woolgar, S. (eds.) (1990), *Representation in Scientific Practice*. Cambridge, MA: MIT Press.
Priestly, J. (1970), *Experiments and Observations on Different Kinds of Air and Other Branches of Natural Philosophy Connected with the Subject in Three Volumes*, vol. 1. New York: Kraus Reprint Co.
Rudwick, M. (1985), *The Great Devonian Controversy*. Chicago: University of Chicago Press.
Shapin, S. and Schaffer, S. (1985), *Leviathan and the Air Pump*. Princeton: Princeton University Press.
Sheehan, W. (1988), *Planets and Perception: Telescopic Views and Interpretations, 1609–1909*. Tucson: The University of Arizona Press.
Stuewer, R. (1985), "Artificial Distintegration and the Cambridge-Vienna Controversy", in P. Achinstein and O. Hannaway (eds.), *Experiment and Hypothesis in Modern Physical Science*. Cambridge, MA: MIT Press, pp. 239–307.
Swijtink, Z. (1987), "The Objectification of Observation: Measurement and Statistical Methods in the Nineteenth Century", in L. Kruger, L. Daston, and M. Heidelberger (eds.), *The Probabilistic Revolution*, vol. 1. Cambridge, MA: MIT Press, pp. 261–285.
Woodward, J. (1989), "Data and Phenomena", *Synthese 79:* 393–472.