THE JOURNAL OF PHILOSOPHY

VOLUME CI, NUMBER 5
MAY 2004

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

Published monthly as of January 1977; typeset and printed by Capital City Press, Montpelier, VT.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY 10027. (212) 866-1742

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.
 ISSN 0022–362X

COMMENTS AND CRITICISM

EVIDENCE, EXPLANATION, AND EXPERIENCE: ON THE HARDER PROBLEM OF CONSCIOUSNESS*

A new type of problem is taking hold in the philosophical debate about consciousness, one that adds to the so-called "hard problem of consciousness."[1] It concerns our epistemic relation to the possible phenomenal states of creatures that are different from us in their physical realization. Ned Block has recently given a thorough account of this "harder problem of consciousness,"[2] and David Papineau, David Chalmers, Joseph Levine, and Thomas Nagel have discussed similar but less substantial versions of this type of problem.[3] It arises for naturalistic phenomenal realists, that is, those who have good but defeasible reasons for believing that consciousness has a scientific nature, and who on the basis of first-person grasp of consciousness believe that phenomenal experiences are instantiations of real properties, not something that can be conceptually reduced away in the manner of analytical functionalism.

Block's harder problem of consciousness (HPC, hereafter) is that naturalistic phenomenal realists face an epistemic tension: they have no conception of a rational ground for believing that other creatures, who do not relevantly share our physical nature, are conscious or not. Even if they assume that the other creature is conscious, they have no conception of a rational ground for believing that consciousness in this creature has a physical basis, even though physicalism is the default view. Furthermore, it is the subjective default that other crea-

* Thanks to the audience at the Annual Meeting of the Danish Philosophical Association, February 2003, especially to Ned Block; and to the audience at a seminar at the Philosophy Program, Research School of Social Sciences, Australian National University. The research for this article was supported by a grant from the Danish Research Council for the Humanities to the research project *Naturalised Mind—Cognisant Nature* (www.namicona.au.dk).

[1] See David Chalmers, *The Conscious Mind* (New York: Oxford, 1996).

[2] "The Harder Problem of Consciousness," this JOURNAL, XCIX, 8 (August 2002): 391–425. Page references are to this article unless otherwise noted.

[3] Papineau, *Thinking about Consciousness* (New York: Oxford, 2002), chapter 7; Chalmers, "Availability: The Cognitive Basis of Experience?" in Block, Owen Flanagan, and Güven Güzeldere, eds., *The Nature of Consciousness* (Cambridge: MIT, 1995), pp. 421–24; Levine, *Purple Haze: The Puzzle of Consciousness* (New York: Oxford, 2001), chapter 3; and Nagel, "The Psychophysical Nexus," in Paul Boghossian and Christopher Peacocke, eds., *New Essays on the A Priori* (New York: Oxford, 2000), pp. 433–71, footnote 21.

0022-362X/04/0105/242–54

tures are not conscious, but the phenomenal realist must leave it an open question whether they are.

Block discusses HPC in terms of a conditional whose antecedent is the conjunction of naturalism, phenomenal realism, and anti-skepticism about other minds (the latter is necessary to distinguish HPC from the other minds problem). The consequent is the epistemic tension. I discuss two aspects of Block's arguments that this conditional is true.

First, I assess the argument to the conclusion that naturalistic phenomenal realists have no conception of a rational ground for belief in disjunctive physicalism. I argue that it is possible to resist the conclusion because, given a plausible notion of inference to the best explanation, the argument begs the question (section I).

Second, I assess the argument that naturalistic phenomenal realism produces an epistemic obstacle to discovering phenomenality in other creatures. This argument ignores the priority of explanation over discovery in inferences to the best explanation (section II). I further explore the relationship between evidence, explanation, and conscious experience: some of the hardness of HPC may stem from doubts about the occurrence of the evidence, or from doubts about the interpretation of the evidence. But the first of these doubts is misplaced, and the second may be overcome, again with help from the notion of inference to the best explanation (section III).

My criticism is not directed at the grounding intuition that there are epistemic obstacles in the study of consciousness; it is pretty clear that we have this intuition. It is directed at the arguments that Block mounts, on the basis of this intuition, to describe the nature and consequences of our current epistemological predicament.

I. THE ARGUMENT FOR META-INACCESSIBILITY OF DISJUNCTIVE PHYSICALISM

Throughout, Block makes use of the science fiction case of Commander Data, who is a merely superficial functional equivalent to us, equivalent to us in respect of folk psychology and what folk psychology entails, but not in other psychological and neuropsychological functions. Importantly, however, Data is unlike us in the physical nature and organization of the control mechanisms of the folk-psychological functions (401–02).

We have good but defeasible grounds for believing that Data is conscious because he acts like us, and we act the way we do in part because we are conscious. The grounds are defeasible because we might find that Data's physical constitution shares none of our neural correlates of consciousness (402–03).

One of Block's central premises is that we have no conception of a rational ground for the belief that Data is or is not conscious. He is stipulated to be physically different from us, so any science of consciousness that is based on us will be powerless to determine whether he is conscious or not. And Block argues further that we cannot base a science of consciousness on him unless we already know he is conscious (I return to this argument in section II). So we have no idea what would be evidence for the belief that Data is conscious or not; Data's consciousness is meta-inaccessible (405).

If the phenomenal realist is a physicalist who thinks that consciousness has a deep, nondisjunctive physical basis, then Data is not conscious because he does not have this basis. But this seems unduly chauvinist. A weaker disjunctivist physicalism allows that Data can be conscious. Disjunctivism says that consciousness is the heterogeneous physical property of having either our or Data's realization of our shared functional state.

Disjunctivism is attractive to the naturalistic phenomenal realist because it allows that we and Data can have overlapping phenomenality with a common, albeit disjunctive, scientific basis. The problem is that, even though it is a default that consciousness has a scientific nature, we have no conception of a rational ground for inferring that disjunctivism is true. It is very odd that someone who believes that consciousness is real and has a scientific nature must also acknowledge that there is no conception of what the evidence for the naturalist position would be (407–08, 413–14).[4]

Block's argument to this conclusion begins by noting that one needs to arrive at disjunctivist physicalism via inference to the best explanation. The identification of phenomenality with the one disjunctive property of our or Data's realizations must contribute to the best explanation of facts about phenomenality, such as why there is phenomenal overlap between us and Data (410–12). This is similar to the way in which the identification of water with H_2O contributes to the best explanation of facts about water, such as that it expands when freezing (409–10). Of course, this only makes sense if it is assumed that Data is conscious: there is no best explanation of the phenomenal overlap, if he is not conscious (411).

There are now two arguments against arguing via inference to the best explanation to the conclusion that consciousness is a heterogeneous physical disjunction.

[4] Block also considers another physicalist doctrine, superficialism (412), which I will ignore here.

The argument about fundamental difference. The first argument in effect begins by stipulating that our and Data's physical realizations are fundamentally different in a sense that makes us balk at a disjunctive explanation of facts about phenomenality. When it comes to two fundamentally different properties, distinct explanations are better than one disjunctive explanation. Block illustrates the point by saying that we could, in theory, explain solid-like formation in terms of crystallization or continuous hardening into an amorphous substance (which would make glass a solid). But we reject the disjunctive explanation and prefer two distinct explanations because the properties of crystallization and continuous hardening into an amorphous substance are fundamentally different. Block does not specify the basis for judgments of fundamental difference, he just notes that we have a vague notion of it and then proceeds to the above stipulation (411).

Block's argument leads up to the conclusion that disjunctivism is meta-inaccessible: that we have no conception of a ground for rational belief in disjunctivism. If our and Data's realizations are fundamentally different, then there can be no argument via inference to the best explanation to a physical disjunction (though disjunctivism may of course still be true).

Nevertheless, this use of the notion of fundamental property difference is questionable. To begin with, it seems doubtful whether a merely intuitive notion of fundamental difference should guide our grounds for rational belief. Intuitively (that is, in the absence of knowledge of the scientific truth), diamonds and graphite are fundamentally different, but they are surely fundamentally similar, since science tells us they are different crystallizations of carbon.

Then there is the case of solid-like formation. It is my understanding that water can freeze into a glassy, amorphous solid when it is cooled rapidly,[5] just as in the case of window glass, which is sand that is melted and rapidly cooled. So here we have solid-like formation—freezing into glassy ice—that is explained in terms of continuous hardening, not crystallization. This opens the possibility of a disjunctive explanation, at least by the indicators employed by Block: glassy ice is *not* a supercooled liquid (glassy solids are more viscous, though still metastable, counterparts of supercooled liquids[6]), and its formation *is* a kind of freezing. But then a reasonable case can be made

[5] Philip Ball, *H_2O: A Biography of Water* (London: Weidenfeld and Nicholson, 1999), p. 187. Block notes this phenomenon as well, and says we do not regard it as freezing (411).

[6] Osamu Mishima and H. Eugene Stanley, "The Relationship between Liquid, Supercooled and Glassy Water," *Nature*, CCCXCVI (1998): 329–35.

that we have an explanation of solid-like formation in terms of a disjunction of intuitively fundamentally different properties, namely, continuous hardening-or-crystallization. So intuitive fundamental difference is not in itself sufficient to close off the feasibility of disjunctive explanation.[7]

The point is not that the discovery of glassy ice straight off should make us accept the disjunctive explanation; it is that the discovery makes it reasonable to suggest that the explanation is disjunctive, in spite of the intuitive fundamental difference. An epistemically defensible notion of fundamental property difference may therefore depend on empirical work and on explanation, and is thus not so easily up for uncontroversial stipulation. In particular, it seems plausible that judgments of fundamental property difference and similarity are, at least in part, based on how well assumptions of fundamental similarity or difference contribute to overall explanatory integration.

The underlying thought is that we would not rationally care much about fundamental similarity and difference unless it had an impact on the overall explanatory picture. If we achieve the best explanatory integration of a wide range of phenomena by one grouping of properties rather than another, then we go with the first. Allowing the disjunctive explanation of solid-like formation may be messier, and less conducive to overall integration than going with the two distinct explanations (it may be less messy in the overall scientific picture to count glass as a liquid). Conversely, once we find out about the glassy solid state of water we might well decide that it is the disjunctive picture that is less messy. It is hard to think of an alternative basis for judgments of fundamental property difference, short of some kind of aesthetic preference that we can hardly allow to inform rationally these epistemic issues.

On this view, the stipulation that our and Data's realizations are fundamentally different is not independent of issues concerning best explanation. Rational judgments of fundamental property difference come *after* we evaluate which explanations are best; we cannot avail ourselves of them first and then use them to tell us which explanations are best. But then the best way to make sense of the stipulation presupposes that there is no argument from inference to best explanation to a physical disjunction. This begs the question, for that is the conclusion at which Block is aiming.

[7] We could take this to show that *solid-like formation* is a merely nominal category, like *jade*. But this would beg the question since it in effect says that only nominal categories can have disjunctive explanations.

Objection: it is possible to obtain scientifically informed views on fundamental property difference on the basis of all *other* explanations than explanations of consciousness, and then to pass judgment on the fundamental difference of our and Data's realizations. *Reply*: this is unduly conservative: even if they were judged different on the basis of other explanations, the best-making qualities of explanations of consciousness may override this judgment (just as the best-making qualities of the explanation of solid-like formation may override previous judgments, once we learn about glassy solids).

Objection: Block's argument explicitly concerns our epistemic predicament in a pre-explanatory situation (406, 424); so it begs the question against Block to invoke what we may say about property difference once explanations become available. *Reply*: the argument is not that HPC will disappear once explanations of consciousness become available, it is rather that Block faces a dilemma: be irrationally guided by pre-explanatory judgments about fundamental property difference or beg the question by presupposing that there is no best disjunctive explanation.

The argument about bad and good questions. Block's second argument against arguing via inference to the best explanation to the conclusion that consciousness is a heterogeneous physical disjunction concerns the neatness of best explanations. Considerations of neatness say that best explanations should solve more explanatory puzzles, and not give rise to bad explanatory puzzles. The theoretical identification of water and H_2O, and heat in gases and mean molecular kinetic energy, prevent bad questions about why it is that water is *correlated* with H_2O and why it is that heat is *correlated* with mean molecular energy from arising; there is no explanation of why things correlate, if they are identical. According to Block, this is part of the reason why we accept these identities.

Now consider the question of why we overlap with Data in this or that phenomenal respect. This question does not arise if we accept disjunctivism. If there is phenomenal overlap, then this just *is* our having the disjunctive property of our or Data's realization—and identities do not have explanations. Block now claims that the question does not seem bad, and therefore should not be ruled out, so, accordingly, we should not accept the theoretical identification. (This is not evidence that Data is not conscious since, for the phenomenal realist, it is still an open question whether he in fact is.) If inference to the best explanation cannot do the job, then there seems no source of a conception of a rational ground for belief in disjunctivism.

We have to ask, however, what makes us deem that certain why-questions are good and others bad. In an argument about the epistemic viability of physicalism, we cannot let ourselves be rationally guided merely by how questions *seem*; that is too arbitrary. Moreover, it seems plausible that such judgments can be wrong: what we once deemed a good question may later transpire as a bad question, and vice versa. My view is that it is explanation that can make us change our minds about the quality of why-questions. We say that the question about why it is that water is correlated with H_2O is bad because the best explanation of facts about water involves acceptance of the theoretical identification of water and H_2O. If we had no such explanation it is very plausible that it would be deemed a good question, at least by some (such as staunch phlogiston-believers who are just coming to terms with Lavoisier's notion of *oxygen*).

Some explanations rule out some questions as bad and other explanations rule out other questions. But it is only after we have the overall explanatory picture that we can make an informed judgment about which questions were good and which bad. What this shows is that the ruling out of bad questions, in Block's sense, is not an independent best-maker for explanations, and thus should not in itself be used to plead acceptance or rejection of theoretical identities. But then, under the assumption that there is as yet no explanation, arguments based on these notions are not available, unless the question is begged in favor of there not being a best disjunctive explanation.

Objection: What you argue is that as long as there is no explanation of consciousness, there will be no resolution of these problems, but is that not precisely the situation Block describes—namely, that in our current epistemic situation the questions about physicalism about consciousness are meta-inaccessible? *Reply*: Block is arguing *for* the claim that physicalism is meta-inaccessible; it is not a premise in his argument. My argument is that this argument does not go through without begging the question. The arguments in this section do not concern the *premise* that we have no conception of a rational ground for belief that Data is conscious or not.

II. EXPLANATION AND DISCOVERY

Block's premise about the meta-inaccessibility of Data's consciousness plays an important role in the overall argument. The argument is that any science of consciousness that can generalize to other creatures must be based in part on them in the first place, but that this cannot be done unless we already have discovered whether they are conscious or not (406–07). The premise about Data prevents this discovery. So, as things stand, there will be no generalizable science of consciousness.

Behind this argument is the following epistemic principle about the priority of discovery over explanation:

(DE) There can be no explanation of a phenomenon unless it is already discovered that it occurs.

I think DE gets some of its intuitive plausibility from the following metaphysical principle:

(OE) There can be no explanation of a phenomenon that does not occur.

For example, there can be no explanation of the occurrence of global peace if it does not occur (though of course there can be explanations of the nonoccurrence of a phenomenon, like the nonoccurrence of global peace). OE is weaker and less controversial than DE because it does not require any discovery of the occurrence of the phenomenon; it is consistent with a range of epistemic states vis-à-vis the phenomenon, and with discovering its occurrence after having explained it.

In his influential discussion of inference to the best explanation, Peter Lipton notes that the point of explaining is not just to explain *why* something is the case; in the shape of inference to the best explanation it is also our primary tool for discovering *what* is the case.[8] One of Lipton's examples illustrates this. Consider the red shift observed in the light spectrum of some stars. This was taken to be evidence of some unknown, undiscovered phenomenon. Various hypotheses can be proposed to explain the evidence, and the one that best explains the evidence is the one we come to believe, in this case that the star recedes with a certain velocity. So we used best explanation to discover something about the world. The principle is:

(ED) Sometimes best explanation can lead to discovery.

This principle does not contradict DE. DE says that the explanandum phenomena must already be discovered. In the recession example, the explanandum phenomenon is the red shift, which *is* discovered. The discovery that is relevant to ED concerns the explanans, that is, the recession. The question is, however, whether ED will help circumvent the role that DE plays in Block's argument. Can explanation of various kinds of phenomena lead to discovery of consciousness in other creatures?

What is the best explanation of the evidence of folk-psychological

[8] *Inference to the Best Explanation* (New York: Routledge, 1991), pp. 57–58, 68–69.

functionality in Data? We would normally reason from like effects to like causes and say the explanation is the occurrence of phenomenal states. But this inference is not sound here because we know that Data's realization is different from ours, and this defeats the argument by analogy.

Consider instead the possibility of inferences to hypotheses that explain the organizing principles of the physical properties of Data's silicon brain and how they give rise to the folk-psychological functionality. One of these hypotheses may explain some of this evidence partly in terms that clearly identify the phenomenal property of pain with a particular pattern of silicon brain activity. Another may explain it partly in terms of the identity of a nonconscious, merely cognitive property and that particular pattern of silicon brain activity. We should accept the hypothesis that best explains the evidence. If this hypothesis is the one that identifies phenomenal properties with silicon brain properties, then that is the one we should believe. That is, best explanation can lead to the discovery of phenomenality.

Objection: this discussion of HPC presupposes that we have a conception of explanations of how a physical property can be identical to a phenomenal property but HPC is explicitly set in a pre-explanatory context (406, 424). *Reply*: this is irrelevant to the issue at hand, which concerns an argument about the relative priority of discovery and explanation. It begs the question in favor of the thesis that discovery must precede explanation if we cannot assume for the sake of argument that explanations are available.

Objection: HPC is not a principled problem, it does not attempt to say that we *could* not have a conception of a rational ground for belief in Data's being conscious or not. Rather, it is contingent on the fact that we have not yet managed to explain how a physical property can be identical to a phenomenal property, and thus it merely argues that our current epistemic predicament about consciousness is of a certain nature and has its origin in naturalistic phenomenal realism. It is no criticism of this view to note that HPC is resolved once we arrive at such explanations. *Reply*: it is one thing to note our strong intuition that, if confronted with Data, we would have no conception of a rational ground for belief that he is conscious or not. I do not deny that we have this intuition. It is another thing to *argue* that we need to discover phenomenality before we can have a generalizable science of consciousness, or indeed that disjunctivism is not epistemically viable. Perhaps our intuitions support these conclusions (though I am not sure about that), but it does not follow that the arguments leading up to the conclusions are convincing. My criticism so far is that these arguments are not convincing. So we are left with the intuitions.

III. DOUBTS ABOUT THE OCCURRENCE AND INTERPRETATION OF THE EVIDENCE

The argument in the preceding section establishes that explanation can lead to the discovery of consciousness in other creatures, and thus we can circumvent Block's claim that there cannot be a science of consciousness for creatures like Data. It is, however, possible to resist this conclusion by changing the context of doubt so that it no longer concerns what phenomena we have and have not discovered, but instead concerns our justification for believing in the occurrence of the evidence. This kind of doubt may be the source of the hardness of the harder problem, and it does indeed lead to a very hard problem, but the underlying doubt is misplaced. Another kind of doubt, about the interpretation of the evidence, is not misplaced but can be dealt with by once again employing properties of inference to the best explanation.

Doubts about the occurrence of the evidence. Here is how the doubt arises. We can grant that best explanation can lead to discovery, but only if there are no doubts about the occurrence of the evidence for the explanation. For example, explanation of a red shift can lead to discovery of recession, but only if there are no doubts about the occurrence of the red shift (was it a genuine red shift, or an artifact of our measuring devices?). Similarly, the doubt goes, "silicon brain science" explanations of the evidence concerning the organizing principles of the physical properties of Data's silicon brain and how they give rise to the folk-psychological functionality can lead to discovery of consciousness, but only if there are no doubts about the occurrence of the evidence. But there are such doubts, for perhaps the organizing principles are designed to mimic genuine folk-psychological functionality and are therefore no more evidence of phenomenality than a tape-recording of fabricated introspective reports.

We cannot respond by claiming that if there *is* a best explanation in terms of phenomenality then there are no grounds for doubting the evidence. This kind of argument leads to vicious circularity in cases where the phenomena that are being explained are themselves essential parts of the evidence for the explanation. If someone doubts the occurrence of the red shift in the light spectrum of a star, then we cannot justify belief in its occurrence with reference to the best explanation (that is, in terms of recession), since the red shift is evidence for the recession. Similarly, if someone doubts the occurrence of the evidence concerning the organizing principles of Data's brain and the functionality, then we cannot justify belief in its occurrence with reference to a best explanation (that is, in terms of phenomenality), since it is an essential part of the evidence for that explanation.

This is a general feature of scientific explanation, and in particular of inference to the best explanation. Some explanations are self-evidencing, that is, the phenomena they explain are themselves part of the evidence for the explanation. There is circularity here, but it is benign. It turns vicious, however, when the aim is to use best explanation to establish the occurrence of the evidence.[9]

Hence, in the context of doubts about the occurrence of the evidence, it seems there is no way of forming a conception of a ground for rational—that is, noncircular—belief in Data's being conscious or not, even if there are available explanations that identify phenomenal properties with some physical properties of his brain! In this version, the problem posed by Block's premise about the meta-inaccessibility of Data's phenomenality is therefore very hard. In contrast to the hard problem, it will not immediately go away once we have a conception of what explanations of phenomenality will be like.

Doubts about the interpretation of the evidence. Luckily, this is not a good way to characterize the doubt that one may have about the evidence. For in a very real sense there is no doubt about the occurrence of the evidence. Data's folk-psychological behavior, and the silicon circuitry in his brain, is in principle there for all to observe. We have no reason to introduce some fault in our methods for acquiring this kind of evidence.

I think there is a better way to capture the relevant doubt about the evidence, but that the problem posed by this doubt is solvable such that we can have a plausible conception of a rational ground for belief, also post explanation. It is possible to view the problem as concerning the interpretation of the evidence rather than the occurrence of the evidence. Sometimes people disagree about the acceptability of various hypotheses because they disagree about the standards of interpretation of the evidence, not because they disagree about whether the evidence confirms the hypothesis. Such disputes seem intractable because any further evidence that gets drawn in to settle the matter is itself subject to the different standards of explanation (which is why this kind of dispute also plays a role in Thomas Kuhn's notion of incommensurability).

By once again employing the resources of inference to the best explanation it is nevertheless possible to decide such disputes.[10] The

[9] Carl G. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press, 1965), pp. 370–72; Lipton, pp. 26–27.

[10] Cf. Lipton, pp. 71–72.

standards of interpretation are partly driven by theories, and if the explanation of the evidence is much better on one interpretation of it than on another, then this vindicates the theory behind the better explanation. For example, we can imagine two different interpretations of the red shift evidence: one in terms of a largely modern astronomical theory, and one in terms of a terracentric, epicycle theory. On the first, it is evidence for and is explained by recession, on the second it is evidence for and is explained by, let us say, some new kind of epicycle. No new astronomical evidence can decide between the hypotheses, but we can nevertheless decide between them because the modern theory allows a much better explanation than the terracentric theory. So the dispute is not intractable, if we have at least one good enough explanation, and if there are independent standards of best-ness of explanation (in Kuhnian cases of incommensurability we may not have shared best-ness standards either).

I think a good part of the epistemic worries about naturalistic phenomenal realism can plausibly be sourced to this kind of doubt about the evidence. Anything a proponent of phenomenality in Data interprets as evidence of phenomenality (or of a particular phenomenal state), the doubter can subject to interpretation under the theory that there is no phenomenality, only functional equivalence (or that it is a different phenomenal state), and vice versa. This explains why we have no conception of what kind of evidence could decide the matter.

Assume now that two different silicon science hypotheses about Data's brain are offered to explain what goes on when his silicon brain produces his apparent introspective reports of, for example, "being in pain." The first hypothesis involves good, well-integrated explanations that show how it can be true that the phenomenal property of pain is identical to particular physical properties of silicon brain activity. The second involves bad, poorly integrated explanations that show how some other phenomenal state such as the experience of jealousy, or some nonconscious, merely cognitive state, is identical to particular physical properties of silicon brain activity.

If we should choose the best explanation, then we are given reason to believe that Data's reports of "being in pain" are reports of that phenomenal state and, further, that there is phenomenal overlap with us when we are in pain. If the second hypothesis had been the better one, then that would be reason to believe that the reports are not of the phenomenal state of pain, and that there is no such overlap.

Hence, if we construe the problem as one concerning the interpretation of the evidence, then inference to the best explanation can again come to the rescue.[11]

Objection: this gets us nowhere because a best explanation must among other things explain the phenomena, and we have no conception of a rational ground for believing that one explanation of phenomenality is better than another (cf. 426). *Reply*: this misconstrues the debate. What is true is that we have no conception of what a scientific explanation of phenomenality may be like: of how physical properties could be identical to phenomenal properties, or how subjective concepts and objective concepts could pick out the same property. This is what remains of the hard problem of consciousness, according to the conceptual dualism espoused by Block (395–98). If we did have such a conception, then we could employ our general notion of inference to the best explanation to reach a decision on Data's being conscious or not; naturalistic phenomenal realism does not throw up a further, separate issue about conceptions of rational grounds for belief.

IV. CONCLUDING REMARKS

The question is whether naturalistic phenomenal realism produces some epistemic problems for the science of consciousness, on top of the problem that it is fiendishly difficult to explain consciousness, and that we currently have no real idea how to go about it. The answer is that consciousness is a singularity because it is so evidentially insulated, and that this makes things difficult but that this does not put the science of consciousness in a league of its own. The epistemic problems associated with naturalistic phenomenal realism are tractable, using the normal resources of scientific practice, in particular as provided by the notion of inference to the best explanation. The trouble with seeing this is that debate of the resources of inference to the best explanation works best on the presupposition that there *are* explanations to infer to, and we resist making that presupposition in the case of consciousness. But then it is no wonder epistemic worries crop up, because the relevant epistemic questions are best assessed in terms of properties of inference to the best explanation.

JAKOB HOHWY

University of Aarhus

[11] This type of strategy is further explored, in the context of a concrete neuroscientific example, in Jakob Hohwy and Chris Frith, "Can Neuroscience Explain Consciousness?" forthcoming in *Journal of Consciousness Studies* (June–July 2004).

THE JOURNAL OF PHILOSOPHY
VOLUME CI, NO. 5, MAY 2004

LIKELIHOOD, MODEL SELECTION, AND THE DUHEM-QUINE PROBLEM*

The Duhem-Quine problem is usually formulated *deductively* with a choice described *dichotomously*: When the conjunction of a hypothesis (H) and an auxiliary assumption (A) entails an observational prediction (O) that fails to come true, should one reject H or reject A? A more general formulation would be to ask what one should say when the conjunction (H & A) confers some probability on O, and instead of considering the two choices just mentioned, the problem would be to evaluate judgments that are a matter of degree. For example, if the observational outcome disconfirms the conjunction (H & A), what determines whether and how much each conjunct is disconfirmed? Indeed, the negative cast of this question can be discarded by generalizing further: How does the disconfirmation or confirmation of the conjunction affect the disconfirmation or confirmation of the conjuncts?[1] The epistemological holism associated with Pierre Duhem[2] and W.V. Quine[3] denies that evidence bearing on (H & A) can have an impact on H that differs from the impact it has on A. This holism can take two forms. *Nondistributive holism* asserts that only whole conjunctions are confirmed and disconfirmed, never their constituent conjuncts; *distributive holism* concedes that evidence

* My thanks to Julian Barbour, Martin Barrett, Richard Creath, John Earman, Ellery Eells, Branden Fitelson, Malcolm Forster, Michael Friedman, Alan Hájek, Daniel Hausman, Margaret Moore, John Norton, Michael Stölzner, Peter Turney, and the editors of this JOURNAL for useful suggestions.

[1] If the original Duhem-Quine problem concerns a question about acceptance/rejection, how is the problem described here concerning confirmation/disconfirmation related to that original problem? If decisions about acceptance and rejection need to include an evaluation of the evidence at hand, the second problem is part of the first.

[2] *The Aim and Structure of Physical Theory* (Princeton: University Press, 1954).

[3] "Two Dogmas of Empiricism," in *From a Logical Point of View* (Cambridge: Harvard, 1953), pp. 20–46, and *Philosophy of Logic* (Englewood Cliffs, NJ: Prentice-Hall, 1970).

0022-362X/04/0105/221–41

bearing on the conjunction can have an impact on a conjunct, but insists that the effect on one conjunct must be the same as the effect on the other.[4] Holists grant that hypotheses and auxiliary assumptions are often treated differently when predictions fail, but claim that it is *nonevidential considerations,* such as simplicity or conservatism, that does the work.[5] To refute holism, the challenge is to show how *the evidence* can have an effect on hypotheses that differs from its effect on auxiliary assumptions.

Previous attempts to bring probabilistic tools to bear on the Duhem-Quine problem have mainly been Bayesian.[6] It is intrinsic to this approach that one must discuss $\Pr(H|A)$, $\Pr(A|H)$, and the probability of the observations conditional not just on (H & A) but on (H & not$A$) and on (not$H$ & A).[7] It is not essential to assign point values

[4] Distributive holists may assert that the effects on H and A are *qualitatively* the same (that is, that both are confirmed or both are disconfirmed) or, more ambitiously, that the effects are *quantitatively* the same (that is, that the degree of confirmation of H is identical with the degree of confirmation of A). For more on this taxonomy of holisms, see my "Quine's Two Dogmas," *Proceedings of the Aristotelian Society,* LXXIV (2000): 237–80.

[5] I do not concede that simplicity *is* always an extra-evidential consideration; the point here is that this is what holists happen to believe. For discussion, see my *Reconstructing the Past: Parsimony, Evolution, and Inference* (Cambridge: MIT, 1988), and my "Instrumentalism, Parsimony, and the Akaike Framework," *Philosophy of Science,* LXIX (2002): S112–S123.

[6] See, for example, J. Dorling, "Bayesian Personalism, the Methodology of Scientific Research Programs, and Duhem's Problem," *Studies in the History and Philosophy of Science,* X (1979): 177–87, Colin Howson and Peter Urbach, *Scientific Reasoning: The Bayesian Approach,* (La Salle, IL: Open Court, 1989), John Earman, *Bayes or Bust?* (Cambridge: MIT, 1992), and Michael Strevens, "A Bayesian Treatment of Auxiliary Hypotheses," *British Journal for the Philosophy of Science,* LII (2001): 515–37. For a nonBayesian treatment, see Deborah Mayo's *Error and the Growth of Experimental Knowledge* (Chicago: University Press, 1996), which analyzes the Duhem-Quine problem within the context of frequentist statistics.

[7] As a simple example of a Bayesian analysis, let us define the degree of confirmation that X receives from Y, $c[X,Y]$, as the ratio $\Pr(X|Y)/\Pr(X)$. So defined, $c[X,Y] > 1$ when Y positively confirms X and $c[X,Y] < 1$ when Y disconfirms X. By Bayes's theorem, this ratio equals $\Pr(Y|X)/\Pr(Y)$; thus, $c[H, \text{not}O] > c[A, \text{not}O]$ if and only if $\Pr(\text{not}O|H) > \Pr(\text{not}O|A)$, which expands to

$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(A\,|H) + \Pr(\text{not}O|\, H \,\&\, \text{not}A)\Pr(\text{not}A\,|H) >$$
$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(H\,|A) + \Pr(\text{not}O|\, \text{not}H \,\&\, A)\Pr(\text{not}H\,|A).$$

Note the occurrence of $\Pr(A|H)$ and of $\Pr(H|A)$ in this expression. If we assume that H and A are probabilistically independent, the inequality reduces to

$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(A) + \Pr(\text{not}O|\, H \,\&\, \text{not}A)\Pr(\text{not}A) >$$
$$\Pr(\text{not}O\,|H \,\&\, A)\Pr(H) + \Pr(\text{not}O|\, \text{not}H \,\&\, A)\Pr(\text{not}H).$$

Notice the prior probabilities of A and of H. I do not mean to beg questions here about the proper definition of degree of confirmation, on which see Branden Fitelson, "The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity," *Philosophy of Science,* LXVI (1999): S362–S378. The point is just to identify the kinds of quantities that a Bayesian analysis must evaluate.

to these quantities; the formal treatments require only that value ranges or inequalities among these quantities be provided. The problem with this approach is that these quantities are often difficult to interpret *objectively*. What is the probability of Newton's theory if there are seven planets? What is the probability of there being seven planets, if Newton's theory is true? And what is the probability that the orbit of Uranus will have a certain shape, if Newton's theory is false and there are seven planets? It does no good to treat these probabilities as subjective degrees of belief. This is unsatisfactory because a subjective interpretation has the consequence that one's "solution" to the problem lacks normative force—one has offered no reason to think that disconfirmation *should* be assigned more to one conjunct than to the other.[8] In other words, Bayesianism in the context of the Duhem-Quine problem encounters the same limitations that Bayesianism often confronts in other settings.

In what follows I will discuss an example of the Duhem-Quine problem in which $\Pr(H|A)$, $\Pr(A|H)$, and $\Pr(O|\pm H \& \pm A)$ (where H is the hypothesis, A the auxiliary assumptions, and O the observational prediction) can be construed objectively; however, only some of those quantities are relevant to the analysis that I provide. The example involves medical diagnosis. The goal is to test the hypothesis that someone has tuberculosis; the auxiliary assumptions describe the error characteristics of the test procedure. Although it can make sense to talk about the objective probability that someone (randomly drawn from a given population) has tuberculosis and it also can make sense to talk about the objective probability that a test procedure has a certain set of error characteristics, neither of these quantities will enter into the analysis. The analysis proceeds entirely via *likelihoods*; what one needs to consider is just the probability of the observations conditional on four conjunctions of the form $(\pm H \& \pm A)$.[9] It is a special feature of the example that all four of these conjunctions are *simple statistical hypotheses* in the technical sense that each unambigu-

---

Another Bayesian approach would be to compare, not the *change* in probability that O induces in H with the change it induces in A, but the *absolute values* of $\Pr(H|O)$ and $\Pr(A|O)$. Bayes's Theorem entails that $\Pr(H|O) > \Pr(A|O)$ if and only if $\Pr(O|H)\Pr(H) > \Pr(O|A)\Pr(A)$. Prior probabilities occur in this expression, and $\Pr(H|A)$ and $\Pr(A|H)$ are involved as well, since $\Pr(O|H) = \Pr(O|H \& A)\Pr(A|H) + \Pr(O|H \& \text{not}A)\Pr(\text{not}A|H)$ and $\Pr(O|A) = \Pr(O|H \& A)\Pr(H|A) + \Pr(O|\text{not}H \& A)\Pr(\text{not}H|A)$.

[8] See Earman (*op. cit.*).

[9] In what follows, I use "likelihood" and "likely" in this technical sense—the likelihood of $(H \& A)$ with respect to the observation O, is $\Pr(O|H \& A)$, not $\Pr(H \& A|O)$.

ously confers a probability on the observations.[10] After describing how the likelihood concept applies to the example concerning medical diagnosis, I will show how similar patterns can arise in the context of a second inferential framework—that of H. Akaike's criterion of model selection;[11] this time the example will involve phylogenetic inference.

I. AN EXAMPLE AND ITS LIKELIHOOD ANALYSIS

Suppose you want to find out whether someone, whom I will call "Newman," has tuberculosis. To do this, you need to use a tuberculosis test. But what makes something a tuberculosis test? The basic idea is that tuberculosis tests are input-output devices. A sample drawn from the subject (for example, of blood or saliva) is put into a black box, which then goes into one of two states, which we call "positive" and "negative." A good tuberculosis test has small error probabilities. Errors, of course, come in two forms—the test can produce false positives and false negatives. Table 1 represents the four relevant conditional probabilities that characterize a tuberculosis test; each has the form Pr($\pm$ test outcome | $\pm$ tuberculosis). It is important to understand these so-called error probabilities in the right way. For example, e_1 is not the probability that someone with a negative test outcome has tuberculosis; rather, it is the probability that someone with tuberculosis will have a negative test outcome. The relevant probabilities are, so to speak, world-to-device, not device-to-world.[12] It might be less misleading to say that these probabilities measure the test's *sensitivity*; they do not tell you how probable it is that a positive or a negative result is mistaken.[13]

[10] Although Mayo (*op. cit.*) is no friend of likelihood, her point that a good test of a hypothesis H may fail to be a good test of the auxiliary assumption A is congenial to the likelihood analysis I will describe.

[11] "Information Theory as an Extension of the Maximum Likelihood Principle," in B. Petrov and F. Csaki, eds., *Second International Symposium on Information Theory* (Budapest: Akademiai Kiado, 1973), pp. 267–81.

[12] Here I adopt terminology from Hartry Field, who talks about world-to-head and head-to-world reliability. See his "'Narrow' Aspects of Intentionality and the Information-theoretic Approach to Content," in Enrique Villanueva, ed., *Information, Semantics, and Epistemology* (New York: Blackwell, 1990), pp. 102–16.

[13] If our interest in using tuberculosis tests is to find out whether someone probably has tuberculosis, why are error characteristics defined in terms of probabilities of the form Pr($\pm$ test outcome | $\pm$ tuberculosis), rather than in terms of probabilities of the form Pr($\pm$ tuberculosis | $\pm$ test outcome)? The reason is that the latter quantities depend on how rare or common tuberculosis is, but the former do not. It is a curious fact about our universe that probabilities of the form Pr(effect|cause) are often time-translationally invariant, whereas probabilities of the form Pr(cause| effect) rarely are. For discussion, see my "Temporally Oriented Laws," *Synthese*, XCIV (1993): 171–89.

TABLE 1

	S has tuberculosis	*S* does not have tuberculosis
+ test result	$1\text{-}e_1$	e_2
− test result	e_1	$1\text{-}e_2$

How would you go about determining what the error characteristics are of a tuberculosis test? One obvious procedure is to assemble people whom you know have tuberculosis and people whom you know do not, and give each person the test. Suppose that a company in Madison, which is interested in developing a test kit for tuberculosis, does this, with 1000 people in each group. The data the company obtains are given in Table 2. How would you use these frequency data to infer the relevant error probabilities? The standard procedure is to use *maximum likelihood estimation*; you find the estimate that maximizes the probability of the observations. The estimated probabilities for the Madison test procedure are therefore:

(Madison) Pr(+ test result | *S* has tuberculosis) = 997/1000
Pr(+ test result | *S* has no tuberculosis) =2/1000.

Given these probabilities, the *Law of Likelihood* describes how a person's test result should be interpreted. If the test outcome is positive, this result *favors* the hypothesis (H_1) that he or she has tuberculosis over the hypothesis (H_2) that he or she does not.[14] A negative test result has the opposite evidential meaning. The strength of the differential support that a test outcome provides is usually measured by the *likelihood ratio*, which has a value of 997/2 favoring H_1 if the result is positive and a value of 998/3 favoring H_2 if it is negative.[15] These hefty values indicate that the Madison test can be said to be a *good* test for tuberculosis. The test provides a lot of information, regardless of the outcome.

TABLE 2

MADISON	1000 with Tuberculosis	1000 with no Tuberculosis
+ test result	997	2
− test result	3	998

[14] See Ian Hacking, *The Logic of Statistical Inference* (New York: Cambridge, 1965); Anthony Edwards, *Likelihood* (New York: Cambridge, 1972); and Richard Royall, *Statistical Evidence: A Likelihood Paradigm* (Boca Raton, FL: Chapman and Hall, 1997).

[15] Here and in what follows I rely on likelihood ratios to measure strength of evidence. The warning in Fitelson (*op. cit.*) that epistemological conclusions may be sensitive to choice of measure is relevant here.

In addition to the Madison company I have just described, suppose there is a company in Middleton that has been involved in the same project. They also want to develop a tuberculosis test kit, so they also try out their procedure on 1000 people whom they know have tuberculosis and 1000 people whom they know do not. The data they obtain on their kit are given in Table 3. They then use maximum likelihood estimation to estimate the error probabilities of their test:

(Middleton) $\Pr(+ \text{ test result} \mid S \text{ has tuberculosis}) = 990/1000$
$\Pr(+ \text{ test result} \mid S \text{ has no tuberculosis}) = 5/1000.$

Notice that the Middleton device is inferred to have slightly larger error probabilities, both positive and negative, than the Madison test. Still, it is a pretty good test. If someone has a positive result on the Middleton test, the likelihood ratio of the two hypotheses H_1 (S has tuberculosis) and H_2 (S does not have tuberculosis) is 990/5 favoring H_1, and a negative outcome engenders a likelihood ratio of 995/10 favoring H_2.

TABLE 3

MIDDLETON	1000 with Tuberculosis	1000 with no Tuberculosis
+ test result	990	5
− test result	10	995

We now can return to the original problem of finding out whether Newman has tuberculosis. I introduced the two tuberculosis tests to give this problem a Duhemian twist. Duhem emphasized that physical theories do not entail observational predictions all by themselves, but do so only when conjoined with auxiliary assumptions. Duhem's insight is preserved in the example at hand, even though the relationships are probabilistic, not deductive. Suppose we give Newman a tuberculosis test and obtain a positive result. The probability of obtaining that result depends both on whether Newman has tuberculosis and on whether we used a test kit from Madison or one from Middleton. The four probabilities are represented in Table 4.

TABLE 4

		Possible Auxiliary Assumptions		
		A_1: Madison		A_2: Middleton
Hypotheses	H_1: Newman has tuberculosis	997/1000	$>$	990/1000
		$\vee$		$\vee$
	H_2: Newman does not have tuberculosis	2/1000	$<$	5/1000

Notice that there is a *qualitative asymmetry* between what the observational outcome says about the hypotheses H_1 and H_2 and what it says about the auxiliary assumptions A_1 and A_2. Newman's positive test result renders H_1 more likely than H_2, regardless of whether A_1 or A_2 is true. However, whether A_1 is more likely than A_2 depends on which of the hypotheses is true, and this, I am assuming, is something we do not already know. Of course, if we *do* already know whether Newman has tuberculosis, then the observed test result does provide information about whether the test kit came from Madison or from Middleton. However, the information provided is exceedingly modest. If Newman has tuberculosis, the positive test outcome slightly favors A_1 over A_2; the likelihood ratio here is only 997/990. Similarly, if Newman does not have tuberculosis, then the positive result favors A_2 over A_1, with a likelihood ratio of 5/2. On the other hand, if we not only do not know whether Newman has tuberculosis but cannot even assign a probability to this being the case, the test result tells us nothing about the provenance of the test kit.

If the data and the ensuing maximum likelihood estimates of error probabilities from either Madison or Middleton had been different, it could easily have turned out that there is no *qualitative* asymmetry between the observation's impact on the hypotheses and its impact on the auxiliary assumptions. That is, it is possible for the observation to provide information about both. However, this does not mean that the *amount* of information provided must be the same; there still can be a *quantitative* asymmetry, even if there is no qualitative asymmetry. By changing the probability in the lower-right cell in Table 4, we obtain Table 5. Now Newman's positive test result favors H_1 over H_2, regardless of which auxiliary assumption is true, and it also favors A_1 over A_2, regardless of which hypothesis is true. However, the observation provides much more information about whether Newman has tuberculosis than it does about whether the test kit came from Madison or Middleton. With respect to the hypotheses, the ratio is either 997/2 or 990/1, depending on which test procedure was used. With respect to the auxiliary assumptions, the ratio is either 997/990 or 2/1, depending on whether Newman has tuberculosis. Note that 997/2 and 990/1 are both much bigger than 997/990 and 2/1. In this example, the observation provides more information about the hypotheses than it does about the auxiliary assumptions. Of course, the reverse situation can also obtain and it also is possible for the situation to be perfectly symmetrical; two examples of symmetry will be discussed at the end of the paper.

TABLE 5

		Possible Auxiliary Assumptions		
		A_1: Madison		A_2: Middleton
Hypotheses	H_1: Newman has tuberculosis	997/1000	>	990/1000
		∨		∨
	H_2: Newman does not have tuberculosis	2/1000	>	1/1000

When a conjunction (H & A) makes a prediction that neither conjunct makes on its own, epistemological holism says that it is *never possible* for the outcome to have an evidential significance for H that differs from the significance it has for A. The generality of this thesis means that just one counterexample is enough to refute it. I claim that the example just described performs that function. Let H be the hypothesis that Newman does not have tuberculosis and let A be the hypothesis that one is using the Madison test procedure. The conjunction (H & A) predicts that the test result will be negative in the sense that it confers on that outcome a probability of 998/1000. But suppose the test comes out positive. To see what this outcome means for H and what it means for A, we need to know what the alternatives are to each. With the alternatives as described, the test result can have a bearing on the hypotheses that differs fundamentally from the bearing it has on the auxiliary assumptions. Both qualitative and quantitative asymmetries are possible. Epistemological holism is false.

II. TWEAKING THE EXAMPLE

The example just described, in which you do not know beforehand whether Newman has tuberculosis and also do not know which tuberculosis test kit you are using, is somewhat artificial. Scientists typically know the provenance of the test kits they use as well as their estimated error probabilities. But even in this more realistic setting, there still is room for Duhemian puzzlement. Suppose we *know* we are using the Madison test kit. However, we recognize that the error probabilities associated with this test kit are merely estimates—we have no certainty that the estimated values are exactly right. Again we give the test to Newman and again obtain a positive result. Does that outcome provide information about whether Newman has tuberculosis and does it also provide information about the test procedure's error characteristics? If so, does the outcome provide more information about one of these than it does about the other?

It might seem intuitive to say that Newman's test outcome provides zero information about the error characteristics of the test procedure. After all, Newman is quite unlike the 2000 subjects who were used to calibrate the test; we have no independent knowledge as to whether he has the disease. Of course, if we knew that he *probably* has the disease, or that he *probably* does not, that would tell us whether his positive test result is *probably* a true positive or is *probably* a false positive, and that would lead us to modify slightly our estimates of the test's error characteristics. But suppose we do not know even that. How, then, can the test result provide any information at all about the test's error characteristics?

Newman's test outcome could be a false positive or it could be a true positive. Let us consider these possibilities in turn. If Newman's test result is a true positive, we should add this result to the 2000 individuals already studied and change our estimate of the test's error characteristics to

Pr(+ test result | *S* has tuberculosis) = 998/1001
Pr(+ test result | *S* has no tuberculosis) =2/1000.

On the other hand, if Newman's test result is a false positive, we should revise our estimate of the test's error characteristics as follows:

Pr(+ test result | *S* has tuberculosis) = 997/1000
Pr(+ test result | *S* has no tuberculosis) =3/1001.

Of course, we do not know whether Newman's result is a false positive or a true positive, so we do not know which pair of estimates we should use to characterize the procedure's error characteristics. However, this uncertainty does not prevent us from formulating a pair of *conditional estimates*:

(New Madison) If Newman has tuberculosis, then Pr(+ test result | *S* has tuberculosis) = 998/1001 and Pr(+ test result | *S* has no tuberculosis) = 2/1000.

If Newman does not have tuberculosis, then Pr(+ test result | *S* has tuberculosis) = 997/1000 and Pr(+ test result | *S* has no tuberculosis) = 3/1001.

We now can ask whether there is a difference in likelihood between the old estimates (Madison) or the new, conditional, estimates (New Madison) that were obtained by taking account of Newman's positive test result.

Table 6 summarizes the situation; cell entries represent the probability of Newman's positive test result, conditional on different combinations of hypotheses and auxiliary assumptions. Notice first that there is a *qualitative symmetry* between what the observation says about the hypotheses and what it says about the auxiliary assumptions. Newman's positive test result renders H_1 more likely than H_2, regardless of which auxiliary assumption is true, and the result also favors (New

Madison) over (Madison) regardless of which hypothesis is true.[16] However, there is a *quantitative asymmetry*. The observation is *very* informative about whether Newman has tuberculosis; 997/2 and 998/3 are both large. In contrast, the observation is only modestly informative about the choice between the auxiliary assumptions; the ratios are approximately 998/997 and 3/2, and these are both rather small. Why does Newman's test result have such a negligible impact on the estimates of the test's error characteristics? The reason is that Newman is just one person out of 2001. Had we initially estimated the error probabilities by using just 200 subjects, or 20, or 2, Newman would have mattered more.

TABLE 6

		Possible Auxiliary Assumptions		
		Madison		New Madison
Hypotheses	H_1: Newman has tuberculosis	997/1000	<	998/1001
		∨		∨
	H_2: Newman does not have tuberculosis	2/1000	<	3/1001

It may seem odd that I even consider (New Madison). If estimates of the error characteristics of a test procedure must be based solely on frequency data, then speculations about what our maximum likelihood estimates would be if we knew whether Newman has tuberculosis are irrelevant. This sensible attitude flies in the face of epistemological holism—it entails that Newman's test outcome provides considerable evidence about whether he has tuberculosis and zero information about the error characteristics of the test procedure. If this were correct, there would be a qualitative as well as a quantitative asymmetry. The analysis in which (Madison) and (New Madison) are compared comes close to this result, but does not coincide with it exactly. I argued that the evidence *slightly* favors (New Madison) over (Madison), not that the observation is literally *informationless*. In terms of the larger picture of seeing what is wrong with epistemological holism, this difference does not matter. But in terms of the specifics of likelihood reasoning, it does.

III. SIGNIFICANCE OF THE TWEAKED EXAMPLE

This last example illustrates a very general fact about the calibration of measurement instruments and the validation of test procedures in

[16] (Madison) has a lower likelihood than (New Madison) in each row, since $a/b < (a+1)/(b+1)$, if $0<a<b$.

science. The typical situation is that the error characteristics of a test procedure are first ascertained and then the procedure is applied to new individuals. One usually does not already know whether these new individuals have the condition being tested (otherwise, why apply the test?); indeed, one often does not even know whether the new individuals *probably* have the condition. Many of us have opinions about the approximate frequency of tuberculosis in this or that population; if we were prepared to assume that Newman was drawn at random from such a population, we would be entitled to talk about the prior probability that he has tuberculosis. Many scientific tests are not like this. Galileo gauged the reliability of his telescope by training it on various *terrestrial* objects. He used it to identify the flags on ships coming over the horizon and the inscriptions on distant buildings; in all these cases it was possible to determine *independently* whether the reports were correct.[17] Galileo then looked through his telescope at *Jupiter*; his observations provided strong evidence that Jupiter has moons, but little or no information about the telescope's error characteristics. Understanding this asymmetry does not require that one assign a prior probability to Jupiter's having moons. This is fortunate, since Galileo was in no position to assign an objective prior probability to that proposition.[18]

There is a sense in which the likelihood analysis of the tuberculosis example is nonBayesian, but this is not because likelihood is an idea that Bayesianism abhors. On the contrary—likelihood is a fundamental quantity in Bayes's theorem. What I mean is that the analysis does not use the full-blown resources that Bayesianism assumes are available. First, prior and posterior probabilities play no role. Second, a likelihoodist will be happy to compare the likelihoods of two simple statistical hypotheses (S_1 and S_2), but often is loath to compare the likelihoods of a simple hypothesis (S_1) and its negation (notS_1) when

[17] See Philip Kitcher's *The Advancement of Science: Science without Legend, Objectivity without Illusions* (New York: Oxford, 1993), pp. 228–33, and his "Real Realism: The Galilean Strategy," *Philosophical Review*, CX (2001): 151–98.

[18] Galileo estimated the error characteristics of his telescope by using it in problems that involved relatively small terrestrial distances; he then applied this detection device to an astronomical object that was much farther away. There certainly was room to wonder, at the time, how trustworthy this bold extrapolation was. My point here is not to comment on the legitimacy of Galileo's inference, but to note how often scientists use the protocol I described in connection with the tuberculosis test. The question of how the behavior of the measuring device should be *parameterized* (for example, a single set of error characteristics for sightings of all objects, or two such sets—one for objects that are near and another for objects that are far away), as opposed to the question of how values for parameters should be *estimated*, will be discussed later.

that negation is composite. Suppose notS_1 is equivalent to a disjunction of simple hypotheses (S_2 or S_3 or...or S_n). If so, the likelihood of notS_1 will be a *weighted average* of the likelihoods of S_2, S_3,...S_n, where the weighting term has the form $\Pr(S_i \mid \text{not}S_1)$. This weighting term often lacks an objective interpretation. If Newton's theory is false, what is the probability of each of the theory's specific alternatives? Thus the problem with priors often recurs as a problem for likelihoods.[19] It is a very special property of the tuberculosis example that the two hypotheses considered are *exhaustive* (assuming that Newman exists) and the four conjunctions are simple.[20]

The relations of qualitative and quantitative asymmetry that I have described are purely formal, and therefore do not depend on one's interpretation of probability. Still, the question may be asked of what interpretation of probability I am using when I say that the likelihoods evaluated in the example concerning Newman's tuberculosis are "objective." Clearly, I cannot think of probabilities as subjective degrees of belief. But neither do I wish to endorse the objective interpretations—actual relative frequencies, hypothetical relative frequencies, propensities—now on offer. My preference is the *no-theory theory of probability*, which rejects the need for a reductive analysis of what probability statements mean. Probability is a theoretical quantity. It obeys the axioms of probability and it bears nondeductive inferential relations to observed relative frequencies. Probability, like other theoretical magnitudes, cannot be reduced to observations, nor does it need to be.[21]

This brings us to the question of how general the treatment provided here of Newman's tuberculosis test can be said to be. Does the structure of this quotidian example apply to all situations in which the Duhem-Quine problem arises? There are two types of situation in which it does not; I will describe one of them now and postpone the other until the next section. Scientists sometimes react to the predictive failure of a conjunction by formulating an alternative to

[19] See Richard W. Miller, *Fact and Method* (Princeton: University Press, 1999), and my "Bayesianism: Its Scope and Limits," in Richard Swinburne, ed., *Bayes's Theorem* (New York: Oxford, 2002), pp. 21–38.

[20] I have emphasized how the likelihood approach allows one to avoid considering prior probabilities. However, there is a quantity that the likelihood approach requires one to consider, that Bayesian treatments (of the kind described in footnote 7) do not. When all four conjunctions of the form ($\pm H$ & $\pm A$) are simple statistical hypotheses, the likelihood approach will consider the quantity $\Pr(O \mid \text{not}H \,\&\, \text{not}A)$.

[21] See Isaac Levi and Sidney Morgenbesser, "Belief and Disposition," *American Philosophical Quarterly*, I (1964): 221–32, and Levi, *Gambling with Truth* (New York: Knopf, 1967).

one conjunct without bothering to formulate an alternative to the other. For example, when John Crouch Adams and Jean Joseph Le Verrier tried to account for the fact that the conjunction of Newton's theory and the assumption that there are seven planets generates an inaccurate prediction of Uranus's orbit, they did not try to invent an alternative to Newton's theory.[22] Rather, they set their minds to constructing a specific alternative to the auxiliary assumption, and the result was the prediction and confirmation of a new planet, which we now call Neptune.[23] In this case, the comparison was between two conjunctions—(Newton & seven planets) and (Newton & eight planets)—first using the old observations of Uranus's orbit, and then assembling new observations drawn from pointing telescopes in the right direction.[24] This one-sided response to observational anomaly also occurs when the auxiliary assumptions include propositions of pure mathematics. It is entirely customary for alternative empirical hypotheses to draw on the same body of pure mathematics.[25] In such cases, the observations are used to test (H_1 & A) against (H_2 & A), but are not used to test the auxiliary assumption A against an alternative, since none was formulated in the first place. Cases of this sort involve epistemological asymmetry "by default," so to speak; they therefore differ from the case of Newman and his tuberculosis test in which alternative hypotheses and alternative auxiliary assumptions are both on the table for consideration.

Even so, the likelihood concept throws light on cases in which scientists decline to construct a new theory (or decline to construct a new set of auxiliary assumptions). They often do so because they

[22] Although Newtonian theory does not use the concept of probability, it still is a mistake to think of the theory plus auxiliary assumptions as *deductively entailing* a prediction about what one should observe concerning Uranus's orbit. The reason is that the observation procedures used are subject to error, and these error characteristics need to be modeled probabilistically, just as was true for Newman's tuberculosis test.

[23] W.M. Smart, "John Crouch Adams and the Discovery of Neptune," *Occasional Notes of the Royal Astronomical Society*, XI (1947): 33–88.

[24] Adams and Le Verrier were able to *accommodate* the observed orbit of Uranus within the Newtonian framework by postulating an eighth planet, but did Uranus's orbit *confirm* the hypothesis that there is an eighth planet? It is tempting to answer this question in the negative and to insist that it was only the observation of *Neptune* that provided evidence. The broad epistemological question at issue here is whether using an observation to construct a hypothesis means that the observation fails to provide evidence for the hypothesis. See Christopher Hitchcock and Elliott Sober, "Prediction, Accommodation, and the Problem of Overfitting," *British Journal for the Philosophy of Science*, LV (2004): 1–34, for discussion.

[25] See my "Indispensability and Mathematics," *Philosophical Review*, CII (1993): 35–57.

expect that the new construction would have low likelihood, relative to *all* the evidence. Newtonian theory exhibited excellent fit to lots of other data; it would have been a very tall order to construct an alternative theory that does a better job of handling the data on Uranus while still having high likelihood relative to these other observations. The assumption that there are just seven planets was much less enmeshed with other data sets, so it made sense for Adams and Le Verrier to have held on to Newtonian theory while attempting to revise the auxiliary assumption about the number of planets.[26] As is well known, the same strategy met with failure when applied to the problem of explaining Mercury's orbit. Einstein's general theory of relativity was able to solve the problem precisely because it fit the data that Newton's theory also fit, while fitting the data on Mercury better. The old auxiliary assumption, that there is no planet between Mercury and the Sun, turned out to be right all along, notwithstanding the fact that Le Verrier and others considered the possibility that an as-yet unobserved planet (Vulcan) is perturbing Mercury's orbit.[27]

Must a solution to the Duhem-Quine problem give scientists advice on whether they should formulate an alternative to the hypothesis H or an alternative to the auxiliary assumptions A when the conjunction (H & A) generates a failed prediction? I do not think so. Epistemology does not have the burden of predicting that Uranus's orbit should be handled in one way while Mercury's should be handled in another. It took an Einstein (namely, *the* Einstein) to discover this; there was nothing in the anomalous data and their relation to Newtonian theory that indicated what the facts would turn out to be. It is perhaps more reasonable for philosophy in this instance to remain on one side of the divide between the *context of discovery* and the *context of justification*.[28] The likelihood analysis describes how alternatives should be compared once they are formulated, not whether they are worth constructing in the first place.[29]

[26] This point concerning "enmeshment with other data sets" is an appeal to observational evidence, not to the extra-evidential considerations that holists think are essential.

[27] Whereas the Adams/Le Verrier approach to the orbit of Uranus involved "asymmetry by default," this was not the case with respect to later discussion of the anomalous perihelion of Mercury, in that modifications of the auxiliary assumptions and of Newtonian theory were both developed. See N. Roseveare, *Mercury's Perihelion from Le Verrier to Einstein* (New York: Oxford, 1982), and John Earman and Michel Janssen, "Einstein's Explanation of the Motion of Mercury's Perihelion," in Earman, Janssen, and John D. Norton, eds., *The Attraction of Gravitation: New Studies in the History of General Relativity* (Boston: Birkhäuser, 1993), pp. 129–72.

[28] See Hans Reichenbach, *Experience and Prediction* (Chicago: University Press, 1938).

[29] The distinction between *context of discovery* and *context of justification* is distinct

In discussing Newman's test result, the hypotheses considered were exhaustive (assuming that Newman exists), but the range of alternative auxiliary assumptions was narrowly circumscribed. In both versions of the problem, I assumed that applications of a tuberculosis test are independent and identically distributed; the same pair of error probabilities applies each time someone takes a test and the error probabilities that apply when one person takes the test are independent of what the outcomes happened to be when others did so. But surely this too is a background assumption that is up for grabs; it is not immune from revision in the context of the Duhem-Quine problem. The different auxiliary assumptions we considered disagree about the *values* of two parameters, but they agree on how the problem should be *parameterized*. It is perfectly legitimate to consider alternative *models* of how the test procedure works, where different models parameterize the problem in different ways. This takes us to our next topic.

IV. MODEL SELECTION

Although the likelihood approach is enough to show that epistemological holism is false, I do not claim that it is able to handle all instances of the Duhem-Quine problem. The main limitation concerns the treatment of composite (nonsimple) statistical hypotheses whose likelihoods cannot be interpreted objectively.[30] For example, consider the hypothesis that two physical quantities—say, the temperature and pressure in a closed chamber of gas—are related linearly. Although a specific straight-line hypothesis (for example, $P = 4 + 3T + U$, where U is an error distribution) confers a probability on a given value for pressure, given a value for temperature, it is more puzzling how one should think about the likelihood of the weaker claim that the relationship is linear (that is, that there exist values of a and b such that $P = a + bT + U$). This is because the likelihood of the hypothesis of linearity (LIN) is a weighted average of the likelihoods of all possible straight lines ($L_1, L_2, \ldots$):[31]

$$\Pr(\text{Data} \mid \text{LIN}) = \sum_i \Pr(\text{Data} \mid L_i)\ \Pr(L_i \mid \text{LIN}).$$

If one has no objective basis for saying how probable this or that straight line is, conditional on (LIN), one will not be able to treat the likelihood of (LIN) as an objective quantity.

---

from the distinction between *rules for accepting and rejecting hypotheses* and *rules for saying which hypotheses are better supported by the evidence.* I draw the latter distinction *within* the context of justification.

[30] See M. Forster, "Bayes and Bust: Simplicity as a Problem for a Probabilist's Approach to Confirmation," *British Journal for the Philosophy of Science*, XLVI (1995): 399–429, and my "Bayesianism: Its Scope and Limits" (*op. cit.*).

[31] As an expository convenience, I represent the average likelihood of (LIN) as a discrete summation, not as a continuous integration.

The term "model" is used in the statistics literature on model selection to refer to hypotheses that contain at least one adjustable parameter. The hypothesis of linearity is a model in this sense, but specific straight-line hypotheses are not. Perhaps the most widely used model selection criterion is the Akaike information criterion (AIC), proposed by Akaike (*op. cit.*).[32] AIC is nonBayesian; the goal is not to compute the probability of a model or its likelihood. Rather, AIC aims to provide an estimate of the model's *predictive accuracy*.[33] But how can (LIN), as opposed to a specific straight-line hypothesis, provide a prediction (either accurate or inaccurate) about the gas's temperature when the gas is raised to a particular temperature? The answer is that the model must first be fitted to a set of old data; the parameters a and b are estimated from that data, using the method of maximum likelihood estimation. By substituting these estimates for the adjustable parameters, (LIN) is replaced with the specific straight-line hypothesis L(LIN); this is the specific straight line that renders the old data maximally probable. One then draws a new data set from the chamber of gas and determines how well L(LIN) predicts this new data. The average performance of (LIN) in this prediction task—first being fitted to old data, then seeing how well the fitted model predicts new data—defines the model's predictive accuracy. Estimating the predictive accuracy of models is the goal; the next question is how we should go about attaining that goal. If we have just one data set at hand, how are we to use this evidence to judge how predictively accurate a model is?

An important lesson that scientists absorb from working with models is that making a model too complex will reduce its predictive accuracy. It is easy to get a model to fit the available data by making it sufficiently complex, but the price is often that the fitted model does a poor job predicting new data. This is not a brute fact in the life experience of scientists; rather, Akaike provided a mathematical framework that helps explain why overfitting tends to reduce predictive accuracy. Using some very general assumptions, Akaike proved a result concerning how the predictive accuracy of a model can be estimated. He showed that an unbiased estimate of a model's predictive accuracy can be obtained by looking at two of its properties—how well it fits

[32] See also Y. Sakamoto, M. Ishiguro, and G. Kitagawa, *Akaike Information Criterion Statistics* (New York: Springer, 1986); and Kenneth P. Burnham and David R. Anderson, *Model Selection and Inference: A Practical Information-Theoretic Approach* (New York: Springer, 1998).

[33] See Forster and Sober, "How to Tell When Simpler, More Unified, or Less *Ad Hoc* Theories Will Provide More Accurate Predictions," *British Journal for the Philosophy of Science*, XLV (1994): 1–36.

the data at hand, and how complex it is (where complexity is measured by the number of adjustable parameters the model contains). Akaike's theorem is the basis for AIC, which assigns a score to a model that reflects both its fit-to-data and its simplicity.[34] By comparing the AIC scores of different models, one can estimate which models will make more accurate predictions.

As an example of a model selection problem in which models can be viewed as conjunctions (and so the Duhem-Quine problem can arise), let us consider the task of phylogenetic inference.[35] The goal is to evaluate the plausibility of different phylogenetic trees. Are human beings more closely related to chimps than they are to gorillas, or is the tree topology something different? There are three bifurcating trees that need to be considered—$(HC)G$, $H(CG)$, and $(HG)C$—but none of these confers a probability on the data until some model of the evolutionary process is provided. There are several process models to consider. In the context of molecular evolution, the simplest model is that of T. Jukes and C. Cantor,[36] which assumes that each of the four nucleotides has the same probability per unit time of changing into any of the others. More complex models, such as the one due to M. Kimura,[37] allow different changes to have different probabilities.[38] When a tree topology is conjoined with a process model, its adjustable parameters may be estimated from the data, and the AIC score of the conjunction may then be computed, as depicted in Table 7. Nothing prevents these cell entries from exhibiting the same asymmetries we saw in the likelihood analysis of Newman's tuberculosis test. The data might favor one tree topology over the other, regardless of which process model is used, but fail to provide any robust

[34] Does AIC play into the hands of holism by invoking simplicity? I would say not, in that the justification for AIC depends on empirical assumptions (though ones of great generality). In addition, I will be using AIC only as an example of a model selection criterion; *cross-validation* is another such criterion, and it involves no appeal to simplicity. As it happens, take-one-out cross-validation is asymptotically equivalent with AIC. See M. Stone's "Cross-validatory Choice and Assessment of Statistical Predictions (with Discussion)," *Journal of the Royal Statistical Society*, B XXXVI (1974): 111–47, and his "An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion," *Journal of the Royal Statistical Society*, B XXXIX (1977): 44–47.

[35] See my "The Contest between Likelihood and Parsimony," *Systematic Biology* (2004, forthcoming).

[36] "Evolution of Protein Molecules," in H. Munro, ed., *Mammalian Protein Metabolism* (New York: Academic, 1969), pp. 21–132.

[37] "A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences," *Journal of Molecular Evolution*, XVI (1980): 111–20.

[38] For a survey of the different models now on offer, see Roderic Page and Edward Holmes, *Molecular Evolution: Phylogenetic Approach* (Malden, MA: Blackwell, 1998), pp. 148–62.

indication of which process model is better. It also is possible for the data to provide information about both tree topology and process model while providing more information about one than it does about the other. There is nothing special about AIC in this regard; all model selection criteria can generate both qualitative and quantitative asymmetries. This shows that the solution to the Duhem-Quine problem that I am proposing does not require a commitment to likelihood as the one true way to interpret evidence.

TABLE 7

		Process Models	
		Jukes-Cantor	Kimura
Tree topologies	$(HC)G$	AIC_1	AIC_2
	$H(CG)$	AIC_3	AIC_4

V. CONCLUDING COMMENTS

How is the likelihood analysis described here related to a deductivist formulation of the Duhem-Quine problem in which we recognize that there are three choices (and not just two)—we can reject H_1, reject A_1, or both? If the conjunction (H_1 & A_1) entails O, then $\Pr(\text{not}O \mid H_1 \,\&\, A_1) = 0$. The four likelihoods we need to consider, relative to the observation (not$O$), are shown in Table 8. Of course, the Law of Likelihood does not make recommendations about acceptance and rejection, but merely describes the differential support that the evidence provides. However, if acceptance and rejection require an evaluation of evidence, then the relationship of these four likelihoods is relevant to deciding what to accept and what to reject. If H_2 *dominates* H_1 (that is, if $p_1<p_3$ and $p_2<p_4$), then the evidence favors H_2 over H_1 regardless of which auxiliary assumption is true. If, in addition, A_2 dominates A_1 (that is, if $p_1<p_2$ and $p_3<p_4$), then (H_2 & A_2) is the conjunction with the highest likelihood, and so likelihood considerations point away from both H_1 and A_1. In this circumstance, the likelihood inequalities provide no qualitative asymmetry between the observation's impact on the hypotheses and its impact on the auxiliary assumptions, though a quantitative asymmetry may nonetheless obtain. If dominance holds in one direction but not the other, there is a qualitative asymmetry.

TABLE 8

		Possible Auxiliary Assumptions	
		A_1	A_2
Hypotheses	H_1	$p_1=0$	p_2
	H_2	p_3	p_4

In both the likelihood approach and the model selection approach, all four conjunctions of the form (H_i & A_j) ($i,j = 1,2$) must be considered.[39] This brings out a further difference between these analyses and Quine's holistic epistemology. As noted earlier, the holist recognizes that people manage to decide which conjunct to blame when a conjunction generates a false prediction, but contends that this decision must rely on extra-evidential considerations, such as simplicity or conservatism. The problem is typically formulated in terms of what a person *believes*—if you believe the conjunction (H_1 & A_1) and then find that conjunction refuted by the evidence, which conjunct should you abandon? Quine recommends a policy of "minimum mutilation"—you should impose the smallest change in your web of belief that suffices to restore consistency with the observations, where changing a more "central" belief is regarded as a larger modification than changing a belief that is more "peripheral."[40] It follows that if you believe (H_1 & A_1), there can be no reason for you to abandon *both* conjuncts and embrace (H_2 & A_2), if a more conservative reformation to either (H_1 & A_2) or to (H_2 & A_1) would manage to restore consistency. It is interesting that the Law of Likelihood and model selection criteria place no premium on minimizing change in which propositions you believe. What you happen to believe plays no role at all; in fact, *you* do not enter into the analysis. The question is entirely about the relationship of propositions to data and has nothing to do with people and their affections or mobility. If (H_2 & A_2) has the highest likelihood or the best AIC score, so be it.[41]

The falsity of epistemological holism does not mean that it is *never* true; the point is that it is *often* untrue. Indeed, it is possible to describe a circumstance that can arise within a likelihood framework in which holistic intuitions are vindicated. Suppose that dominance fails, both with respect to the comparison of hypotheses and also with respect to the comparison of auxiliary assumptions. That is, let $p_1 < p_2$, $p_3 > p_4$,

[39] I hope it is clear that my focus on *four* conjunctions is an expository convenience. Both the Law of Likelihood and model selection criteria can address any number of conjunctions, and the asymmetries I have described can arise in that larger comparative context as well.

[40] *Philosophy of Logic*, p. 7; see also "Two Dogmas of Empiricism," p. 44.

[41] Since Quine's principle concerns change in what one believes, whereas the Law of Likelihood and model selection criteria do not provide rules of acceptance, there is no formal incompatibility here. However, when likelihood is placed in the context of a full Bayesian framework with prior probabilities, it can turn out that (H_1 & A_1) is the conjunction with the highest prior probability whereas (H_2 & A_2) is the one whose posterior probability is greatest. The imperative to find the conjunction that differs minimally from the one with the highest prior probability and that is logically consistent with the new evidence is not a principle of Bayesian epistemology.

$p_1<p_3$, and $p_2>p_4$ in Table 8. The cell entries in the table will then have two peaks (at H_1 & A_2 and at H_2 & A_1) and two valleys (at H_1 & A_1 and at H_2 & A_2). If the two peaks have equal likelihood, the observations will not discriminate between them, even though each peak is better supported than the conjunctions that occupy the two valleys. The history of epistemology is peppered with examples of this type. I will mention two; the first is Cartesian, while the second is due to Hans Reichenbach.[42] In both cases, (H_1 & A_1) predicts O, but not$O$ is what you observe:

not$O$: There seems to be a printed page in front of me.
H_1: There is a salami (and not a printed page) in front of me.
H_2: There is a printed page (and not a salami) in front of me.
A_1: My senses are functioning normally.
A_2: An evil demon is causing printed pages to look like salamis, and vice versa.

not$O$: My measurement device indicates that the triangle I have just measured has an angle sum that exceeds 180°.
H_1: Space has zero curvature.
H_2: Space has constant positive curvature.
A_1: There are no universal forces.
A_2: Universal forces are in operation.

Given the observational outcome not$O$, likelihood allows you to discriminate between conjunctions in the same row and conjunctions in the same column, but not between the anti-diagonal entries. The same situation can arise in the context of model selection, if the AIC scores in Table 7 have the pattern of peaks and valleys just described in connection with Table 8.

The likelihood approach views theory evaluation as inherently comparative. The Law of Likelihood describes what an observation says about *two* hypotheses, but is silent on the question of what, if anything, it says about a single hypothesis taken on its own. Likelihoodism elevates that omission to the level of principle; when a hypothesis is tested, it must be tested *against* alternatives.[43] In addition, there is no single alternative that counts as the uniquely correct alternative to consider. There are many specific alternatives, and the Law of Likeli-

[42] *The Philosophy of Space and Time* (New York: Dover, 1958).

[43] An obvious exception arises when a proposition entails an observational prediction that fails to come true. In this instance, there is no need to consider an alternative; *modus tollens* permits one to reject the proposition without further ado. However, in the very common case in which H, or the conjunction (H & A), confers a probability on O that is less than unity, the falsity of O cannot be interpreted until alternatives are considered. Likelihoodism therefore rejects "probabilistic *modus tollens*" (a.k.a. Fisherian significance testing), according to which a hypothesis should be rejected when an event occurs that the hypothesis says was very improbable. For criticisms of probabilistic *modus tollens*, see Hacking, Edwards, and Royall.

hood may be asked for its assessment of all these competitions. This marks an important difference between likelihoodism and Bayesianism. Bayesianism also has its comparative element, but there is just one alternative to a proposition that Bayesians need to consider; this is simply the proposition's negation. The comparative character of likelihood assessments (and of model selection criteria as well) is central to the analysis proposed here of the Duhem-Quine problem: *whether holism is right in what it says about the bearing of evidence on a conjunction crucially depends on what the alternative conjunctions are.* In many scientific contexts, the competing hypotheses and the competing auxiliary assumptions are such that the evidence has an impact on one conjunct that differs from the impact it has on the other. However, there are other discrimination problems in which the evidence cannot penetrate from an assessment of conjunctions to an assessment of conjuncts. Holism has its place, but in the context of scientific inquiry it is, by far, the exception and not the rule.

Quine famously opined that "any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system."[44] Quine's point was not just that it is *logically possible* to continue to believe a proposition in the face of apparently recalcitrant experience, but that *the evidence* does not say that this is a mistake. Decisions about retention or rejection are dictated by nonevidential considerations. I suspect that the allure of this form of holism derives from a hypothetico-deductive view of theory testing. If an observation O confirms a proposition P precisely when P entails O and O turns out to be true, and O disconfirms P precisely when P entails O and O turns out to be false, then epistemological holism is correct when the conjunction (H & A) entails O, but neither conjunct does. In this circumstance the observational outcome confirms or disconfirms the conjunction, but not the constituent conjuncts. However, it is abundantly clear that confirmation and disconfirmation can be mediated by *non*deductive relationships. In the examples discussed here concerning tuberculosis diagnosis and phylogenetic inference, the conjunctions considered were each consistent with the observational outcome without entailing it, but that does not mean that the evidence failed to discriminate among them. Evidence can do more than hypothetico-deductivism imagines. It is that extra power that undermines epistemological holism.

ELLIOTT SOBER

Stanford University
University of Wisconsin/Madison

[44] "Two Dogmas of Empiricism," p. 43.

REVIEW ESSAYS

INCLUSIVE RATIONALITY

Rationality and Freedom. AMARTYA SEN. Cambridge: Harvard University Press, 2002. ix + 736 p. Cloth $39.95, paper $19.95.

Social choice theory is tailor-made for philosophical reflection. The intended applications concern the distribution of benefits to a group of individuals and the design of voting procedures. These topics themselves invite interest from moral and political philosophers. Social choice theory approaches them as problems for rational decision making where the social agent or mechanism confronts a menu of options each of which is the choice of a social state and where the social agent and members of the group evaluate the various options in the menu. The evaluations of the social agent are assumed to be constrained by information concerning the evaluations of the citizens in ways that reflect fairly widespread shared values. The disconcerting result reported by Kenneth Arrow was that these attractive constraints could not all be jointly satisfied. A massive effort to find ways and means to modify the Arrovian constraints so as to avoid the "impossibility" has resulted. These responses have probed into questions of rational choice, epistemology, moral and value theory, and psychology.

Since the pioneering work of Arrow, no one has done more and better to explore the various ramifications of the Arrovian impossibility theorem than Amartya Sen. To take but one very important example, Sen has shown how one might weaken the prohibition endorsed by Arrow on making interpersonal comparisons in an intellectually responsible way so as to finesse Arrovian impossibility and obtain positive results.[1]

Sen's approach to interpersonal comparability is characteristic of his treatment of many other topics. He has sought to adjust the framework of social choice theory in an acceptable manner so that values that seem to be ignored in mainstream economic reflection on welfare economics and social choice can be accommodated. Some of the papers in this volume, including his lecture on the occasion of his being awarded the Nobel Prize in Economics in 1998 (chapter 2)

[1] See *Collective Choice and Social Welfare* (San Francisco: Holden-Day, 1970), chapters 7 and 7*, and *Choice, Welfare, and Measurement* (Cambridge: MIT, 1982), Part III.

0022-362X/04/0105/255–76

and his substantially reedited Arrow Lectures of 1991 (part VI of this volume), are fine specimens of his project. In the Arrow Lectures, Sen rehearses issues that arise from considering freedom as opportunity to choose and freedom as process of choice, their relations to well-being and the preferences broadly construed of agents and argues for the use of a social choice framework comprehensive enough to address these matters.

But the title of this volume is *Rationality and Freedom.* Sen's project is not merely to make social choice theory more sensitive to questions of freedom but to offer a view of rational decision making that has a more inclusive view of rational choice than some of the narrowly specified visions found among economists and decision theorists suggest.

Towards the end of the opening essay of this volume, Sen advocates understanding "rationality as a discipline, not as a favored formula, or as an essentialist doctrine. Rationality includes the use of reasoning to understand and assess goals and values, and it involves the use of these goals and values to make systematic choices" (46).

Rationality is not the property of utilitarianism or other vulgar forms of consequentialism any more than the possession of the aficionados of pure practical reason. The diversity of values that engage human actors and their social agents ought to be accommodated within the confines of an ecumenical vision of rationality. Rationality ought to be understood in a sufficiently capacious sense to allow for legitimate insights from a broad variety of sources and for recognition of the plurality of values that engage human actors and their social agents. Thus rational decision makers should, according to Sen, undertake an evaluation of the options in the "menu" they face by reference to the values they impute to the outcomes of these options or to possible outcomes of implementing these options. But, in contrast to many analysts, Sen insists that outcomes be described in ways that take into account all relevant value commitments of the decision maker. Attention ought not to be restricted to "culmination outcomes" but should be stretched to accommodate when relevant more comprehensive descriptions of outcomes that take into account the process of choice itself. This allows what, at first blush, appears to be a "consequentialist" framework to be applied sensitively to handle the concerns if not the demands of Kant. This distinction is also relevant to Sen's treatment of freedom where he contrasts freedom as opportunity with the "process" aspect of freedom where it becomes relevant whether Mother delivers on the opportunity or Sonny-boy does it himself.

Sen is not alone in protesting the use of narrowly focused ideas of rational choice to rule out as incoherent diverse kinds of value commitments and goals. His distinction lies in his refusal to replace analysis with flaccid posturing. The essays in this volume are a testament to his efforts to refashion the analytical apparatus of accounts of rational decision making and social choice so that dismissing substantive choices and the beliefs and value commitments that motivate them on the grounds of irrationality becomes difficult to do.

My focus in the remainder of this essay shall be on Sen's discussion of rationality or rational decision making that appears in part I chapter 1, part II, and in some of the essays in part III. Sen's examination of freedom and social choice theory reveals the same mix of analytic tough-mindedness, insight, and good-natured generosity of spirit that one finds in Sen's treatment of rational choice. Regretfully, I will not be able to survey his views on these topics here.

Sen stands opposed to two and a half theses widely endorsed by economists and some philosophers: (1) rationality is internal consistency of choice; (2) rationality is self-interest maximization; and (3) rationality is maximization in general (see 19). I say that he is opposed to two and a half theses here because, in contrast to his total rejection of (1) and (2), his dissent from (3) is qualified. If optimization is equated with maximization, he denies that rationality is maximization. According to Sen, rational agents are expected to restrict choice to maximal options among those available to them and maximal options do maximize in a sense (to be explained later) distinct from optimization.

I. CHOICE CONSISTENCY

Recall the "incongruous exchange" described in R.D. Luce and H. Raiffa's *Games and Decisions*:[2]

> DOCTOR: Well, Nurse, that's the evidence. Since I must decide whether or not he is tubercular, I'll diagnose tubercular.
>
> NURSE: But, Doctor, you do not have to decide one way or the other, you can say you are undecided.
>
> DOCTOR: That's true, isn't it? In that case, mark him not tubercular.
>
> NURSE: Please repeat that! (*ibid.*, pp. 288–89).

This is an illustration of an alleged "internal choice inconsistency." We may represent the doctor's choices as values of a choice function C whose domain is a set of finite subsets of some universal set Ω. Any

[2] New York: Wiley, 1957.

such subset represents a hypothetical decision problem where the elements of the subset are the options that would be available for choice in that decision problem (the "menu" of options as Sen puts it). The value of the function is the subset of those hypothetically given options that the agent judges should be admissible for choice in that hypothetical situation. Thus, in the example of the doctor, $C(\{a,b\}) = \{a\}$ and $C(\{a,b,c\}) = \{b\}$. Both the set $\{a,b\}$ and the set $\{a,b,c\}$ are subsets of a set Ω. That set could, in this case, be $\{a,b,c\}$ or some superset of this. In case it is the set of three elements, the choice function has 7 nonempty subsets of Ω as arguments. Let $S \subseteq T$ where S and T are finite, nonempty subsets of Ω. Suppose in addition that $x \in S$. A choice function has property α if and only if whenever $x \in C(T)$, $x \in C(S)$. An agent whose choice function for Ω fails to satisfy this condition automatically violates a condition of internal choice consistency.[3]

Sen argues, and rightly so, that violating such choice patterns is not necessarily a mark of inconsistency in any sense that deserves the charge of irrationality. Perhaps the doctor changed his mind due to some consideration brought to mind by the Nurse's intervention. Without further information about the doctor's motivations, a judgment as to the irrationality of his conduct would be precipitous.

> Statements *A* and *not-A* are contradictory in a way that choosing x from $\{x,y\}$ and y from $\{x,y,z\}$ cannot be. If the latter pair of choices were to entail respectively the statement (1) x is a better alternative than y and (2) y is a better alternative than x, then there would indeed be a contradiction here (assuming that the content of "being better than" requires asymmetry). But these choices do not, *in themselves* entail any such statements. *Given* some ideas as to what the person is trying to do…, we might be able to "interpret" these actions as implied statements (126–27).

As Sen rightly observes, choices do not entail anything and are not inconsistent with one another. The same is true of what is to be chosen in a hypothetical choice situation.

But choice consistency is not intended to be consistency in the sense in which truth-value bearing propositions are consistent. Consistency as a requirement of rationality has often been used in a broader

[3] Sen has discussed the limitations and important uses of notions of choice function, choice consistency, and revealed preference since the late 1960s in *Collective Choice and Social Welfare* and in chapters 1 and 6 of *Choice, Welfare, and Measurement*. To my knowledge, the critique focused on the limitations of *internal* choice consistency that appears in chapter 3 of the book under review constitutes a novel emphasis meriting consideration on its own.

sense than that familiar in deductive logic. "The probability that $h > r$" and "*the probability that* $h \leq r$" can be used to express attitudes that are inconsistent or incoherent in the extended sense favored by Frank Ramsey and Bruno De Finetti. To be committed to such judgments in the same context violates canons of rational probability judgment even though, as both of these authors rightly pointed out, such attitudes are neither true nor false. Perhaps, choosing x from $\{x,y\}$ and choosing y from $\{x,y,z\}$ can be inconsistent in a sense that implies irrationality after all.

Sen's main point nonetheless stands. In order for a choice function C to be subject to rational appraisal, the judgments of admissibility represented by the choice function for the hypothetical situations calling for decision should be constrained by the agent's value commitments and beliefs. The constraints have often been required to satisfy the following conditions:

(i) There must be a "universal set" Ω of candidate options such that the choice function C is a function from nonempty, finite subsets of Ω to finite subsets of Ω such that $C(S) \subseteq S$.

(ii) Decision maker X's value commitments and beliefs determine a value structure $V(\Omega)$ for Ω representable by a weak ordering.[4] X's value commitments and beliefs also determine such a value structure $V(S)$ for every finite, nonempty menu $S \subseteq \Omega$.

(iii) The preference is *menu independent* with respect to Ω (see 172). Given any nonempty subset S of Ω, the value structure $V(S)$ of S when S is the restriction of $V(\Omega)$.[5]

(iv) $C(S)$ is the set of options in the option menu S that are optimal according to $V(S)$. $C(S)$ is thus a *menu independent choice function* (173).

Conditions (ii) and (iv) will be weakened later. For the present illustrative purposes, we shall remain with (i)–(iv) as stated. If conditions (i)–(iv) are met, the choice function must satisfy the requirements of property α mentioned above and the following property β: If $x,y \in S \subseteq T$, $x,y \in C(S)$, then $x \in C(T)$ if and only if $y \in C(T)$. Sen's critique of internal choice consistency emphasizes that whether or not conformity with α and β for specific Ω and C is rationally required or not will depend upon whether the decision maker X's

[4] The weak preference relation xRy (x is as good as y) is reflexive (xRx), transitive ($xRy \& yRz \rightarrow xRz$) and complete ($xRy \vee yRx$). Strict preference xPy holds if and only if $xRy \& {\sim} yRx$. Equipreference xIy holds if and only if $xRy \& yRx$. That is to say, the ordering is not only *complete* but *connected*. $xPy \vee xIy \vee yPx$.

[5] Such preference is "robust" according to Levi, *Hard Choices* (New York: Cambridge, 1986), 6.3.

value commitments and beliefs require satisfaction of (i)–(iv) for the given Ω and C. If they do, X's choices violating α are inconsistent *with those value commitments and beliefs* in a sense that is offensive to reason. However, there is no principle of rationality that entails that X endorses those value commitments and beliefs. And there is no principle of rationality that mandates *internal* choice consistency stipulating that X conform in the choices made to α regardless of X's value commitments and beliefs. As Sen says, without appeal to what the agent is "trying to do," no universal set is specified and no choice function is singled out with respect to which preference and choice function are menu independent.

Sen is entirely right to object to rationality as internal choice consistency. Yet, there is one aspect of his discussion with which I have some difficulty. In marshalling his illuminating counterinstances to internal choice consistency, Sen neglects to indicate how a universal set is constrained by what the agent is "trying to do"—that is, by the agent's value commitments and beliefs. It seems to me that when the agent's value commitments and beliefs are taken into account, the issue does take on a nontrivial significance.

Sen has been interested in the question of "positional choice," in particular, for a long time due in part to its relevance for social choice theory where several voting procedures are positional. He examines several illustrations outside of social choice theory exemplifying the same process. Here is a useful example.

X is presented by his host with a basket of fruit consisting of mango m_1, apple a_1, and apple a_2. X desires to eat a mango but courtesy requires that he decline the only mango in the basket and so he chooses one of the apples.

Had the basket, counter to fact, contained another mango m_2, X would have chosen one of the mangos—perhaps, m_1. The set of potential options Ω includes taking mango m_1, mango m_2, apple a_1, and apple a_2 and refusing any fruit. The menu of options available to X is the set $S = \{m_1, a_1, a_2, \text{refuse}\}$. X's "all things considered" evaluation ranks a_1 and a_2 as best, refusal next, and m_1 at bottom. $C(S) = \{a_1, a_2\}$. But in a case where there are several mangos and several apples, X may choose a mango. $C(T) = C(\{m_1, m_2, a_1, a_2, \text{refuse}) = \{m_1, m_2\}$. Property α has been violated.

Sen points out that there is nothing irrational here. The choice function fails to be menu independent as (iv) requires and the preference fails to be menu independent as (iii) demands. Under the circumstances, rationality does not require conformity with property α. X's value commitments include the injunction never to take the only mango from a basket when others are present. And that injunction does not change when the options are all those in Ω. But the value structure for *S* is not embedded in the value structure for *T*. This menu dependence of the preference is explainable as due to a difference in *X*'s information (and, hence his beliefs) about his options in the two scenarios. Ω has not changed. *X*'s value commitments have not changed. *X*'s information about *X*'s options has changed. In the first scenario, choosing m_1 is known to be equivalent to choosing the only mango. In the second scenario, choosing m_1 is known to be a case of choosing one of several mangos. This change in information alters the change in the evaluation of the options. Hence, the choice function defined over finite nonempty subsets of Ω fails to satisfy α.

It is possible, however, to describe *X*'s judgments of admissibility using a choice function C' of menus carved out of a different universal set Ω' that does, according to *X*'s value commitments and beliefs, satisfy α. It can be argued that the options in the first case are {choosing the only mango, choosing one of several apples, refusal}. Perhaps, in deference to those who insist that there are four options and not just three, we may consider {choosing the only mango—to wit, m_1, choosing one of several apples—to wit, a_1, choosing one of several apples—to wit, a_2, refusal}. In the second case, the options are {choosing one of several mangos, choosing one of several apples, refusal} or the expansion to five options along the lines just indicated. Ω' is a superset of the union of these two sets. Notice that Ω' is different from Ω. *X* intends to do something different in choosing the only mango, namely m_1, than *X* intends in choosing one of several mangos, namely m_1.

It may perhaps be wondered how these two options can be considered part of the same universal set Ω'. To do so is to think it entertainable at least hypothetically that *X* can face a menu consisting of just this pair of options. This seems absurd. But it is surely coherent to consider a scenario where *X* is offered a choice between selecting a mango from a basket containing only one and nothing else and a basket containing two and nothing else. If one keeps in mind that the object of choice is an action of some type or kind and that the universal set is a set of such types of action, there should be nothing problematic here. With Ω' as the universal set and *X*'s value commitments and information, the preference over Ω' is menu independent as is the choice function C' defined over subsets of Ω'. This choice function has property α.[6]

Sen thinks that such redescription strategies trivialize the question

[6] These remarks do not refute theorem 6.3 of Sen, on 163. There the universal set is held fixed. What Sen overlooks is the possibility of adjusting the universal set so as to obtain a menu independence over the preference set.

of choice consistency. But what gets trivialized? No matter what the universal set is over which the choice function is defined, compliance with property α may fail as a matter of fact. Nor is such conformity rationally required for choice functions defined over arbitrary universal sets. It is not required for choice function C defined over Ω. It ought to be required for C' defined over Ω'. That is because *X*'s value commitments and beliefs entail that conditions (i)–(iv) are satisfied. And when conditions (i)–(iv) are satisfied, α ought also to be satisfied. But it is far from trivial that *X* has the value commitments and beliefs that support conditions (i)–(iv) and a choice function satisfying α. What is trivialized?

In the example under discussion, consideration of the value structure over Ω' shows that at least one universal set of potential options and a value structure consonant with *X*'s value commitments and beliefs over that universal set can be constructed such that both menu independence and choice function independence and, hence, choice consistency are required.

Both Ω and Ω' can be used to describe *X*'s choices. But in this example Ω' is better suited to represent *X*'s propensities to deliberate choice. Ω' utilizes descriptions of *X*'s choices relevant to the value commitments of *X*. Any further specification of the options Ω' would be irrelevant to determining whether *X*'s choices conform with *X*'s value commitments and beliefs. If the options are represented as they are in Ω, relevant information is left out. In order to obtain a redescription that produces the information relevant to the evaluation of the options, we need to derive the descriptions in Ω' appealing to *X*'s epistemic state. We can, indeed, do this. But there is nothing trivial about the fact that the specification of the options in Ω' provides the information relevant to promoting *X*'s goals and values whereas the specification of the options in Ω does not.

The moral of the story is this: whether so-called choice consistency conditions ought rationally to be satisfied or not depends upon the universal set relative to which choice functions are defined and the value structure defined over it. Roughly speaking, the specification of the space of potential options over which a choice function is to be defined ought to be one that includes for each option the information relevant to its appraisal according to *X*'s value commitments and beliefs.[7] Call this the condition of *Universal Set Relevancy*. In our toy

[7] Sen offers his counterinstances to α in the context of a study of "choice behavior" rather than in the setting of "normative choice theory" (159 n.2). But a study of choice behavior that does not describe choices in a manner relevant to assessing the rationality of such behavior in terms of the goals and values of the decision maker ought, according to Sen, to be of doubtful usefulness.

example, Ω used by Sen to supply illustrations of failures of choice consistency fails Universal Set Relevancy. Ω' on the other hand does meet the condition. *X*'s choice function conforms with choice consistency requirements relative to Ω'.[8]

As far as Sen's argument goes, therefore, rational agents might be expected to satisfy *relevant* choice consistency conditions even though they need not satisfy *internal* choice consistency conditions. Obeying choice consistency relative to universal sets that satisfy Universal Set Relevancy may prove disconcerting to those who engage in empirical investigations of consumer behavior by seeking to rely on choice behavior exclusively. Unless we are confident that changes in prices and income do not alter the "indifference maps" of individual consumers, a given space of commodity bundles may not qualify as a universal set of potential options.

There are situations where the relevant universal set is generated according to some definite procedure by a core set *S* of options that may or may not be the menu of available options confronting the decision maker. A familiar technique due to John von Neumann and Oskar Morgenstern identifies Ω with the *Mixture Set* $M(S)$ of *value neutral* lotteries using stochastic devices to select an element of *S* to implement. On the supposition that the preference over $M(S)$ in which the preference over *S* is embedded satisfies the von Neumann-Morgenstern axioms, numerical values may be assigned members of the mixture set $M(S)$ that are expectations of the values assigned members of *S*. It is important to recognize, as von Neumann and Morgenstern themselves recognized, that such conditions preclude assigning any value to the process of implementing a chance mechanism to select an element of *S*. If the von Neumann-Morgenstern method for eliciting cardinal utilities is to be deployed, a mixture of pure options is to be evaluated in a way that suppresses any taste or aversion for gambling on that mixture.

Sen rightly objects to requiring as a matter of rationality that choosing a lottery to determine which "pure" option is to be implemented be value neutral in this sense. He points out that there is nothing

[8] Sen seems prepared to recognize the possibility of converting menu dependent preference to menu independent preference by adjusting the universal set in some contexts (178). But he insists that when an agent follows a rule or mode of choice that is menu dependent, this approach is not allowed (178–79). However, I have just constructed a requisite universal set Ω' for menu dependent rules that generates a menu independent preference and menu independent choice function. And I am arguing that the universal set constructed is appropriate for the evaluation of the rationality of such rule following choice. The menu dependence of the rule to be followed does not imply either menu dependent preference or choice function.

irrational about shunning a choice between "pure" options in S in favor of mixtures of such options so as to avoid direct responsibility for the element of S that is implemented. Dr. Chang must choose between giving the only dose of a given medicine available to one of two children in life or death circumstances where one of them has a slightly better chance of survival if treated than the other. He wishes to dodge making the choice and opts instead for the use of a random device even though the expected value of doing so is less than one of the pure options (175–76). There is nothing irrational in Chang's decision—although avoiding tough choices in this way may express value commitments rationalizing some form of moral cowardice.

Where I disagree with Sen is in his contention that behavior like Chang's violates the independence postulate that enforces the value neutrality of mixtures. Chang's behavior is an indication that the lottery he chooses is not a member of the mixture set of neutral mixtures in $M(S)$. Instead of thinking of Chang's set S of pure options as the options of giving the treatment to one of the two candidate patients, the pure options belong to set T which is the set S plus the option of using the random device. The mixture set $M(T)$ is *not* a set of options available to Chang any more than the set $M(S)$ is. Chang can evaluate the options in $M(T)$ on the belief contravening supposition that they are available and on the additional belief contravening supposition of value neutrality. Notice that the option of using the random device to select a patient appears in the list twice: once as a pure option in T that is not a value neutral mixture and the other as a value neutral mixture. Under the suppositions as specified, the preference ranking over $M(T)$ obeys the independence postulate. The item that appears twice is described differently and assigned different values relative to the two descriptions. When described as a pure option of avoiding a direct decision as to who is to be treated, the value imputed is greater than the value of treating the patient most likely to be cured. Described as a value neutral mixed option, the value imputed is less than the value of treating the most promising patient. All of this conforms to the independence postulate and expected utility.

Such hypothetical reflection could bring clarity to Chang's deliberation. If Chang is conscientious, he might want to ponder whether the premium he is placing on avoiding the direct choice of the best pure option is worth the reduction in the expected benefit to the patient who would have otherwise received the treatment for sure. To ask this question presupposes the requirement of the independence postulate on the evaluation of value neutral mixtures as a condition on rational preference. But it does not mandate that rational agents evaluate the mixed options they face as options in a value neutral way.

Sen is aware of views of this kind. He declares that they rob expected utility theory of "much of its operational content" (176). The allegation may be true according to a very narrow reading of operational content. Choice functions characterize an agent's judgments as to what he would choose or judge admissible for choice in various hypothetical and, indeed, typically belief contravening circumstances. This is always the case. There is no diminution of operational content resulting from following the standard practice in Chang's case.

Nor does the account just proposed of the Dr. Chang example support his additional claim that the expected utility axioms are "trivially fulfilled" (176). Dr. Chang could evaluate a mixed option on the supposition that it is value neutral in just the same way he evaluates the lottery to decide which patient to treat. He would indeed violate the dictates of the independence postulates. Fulfilling the requirements of the axioms is not trivial. The issue is not whether Chang could violate conditions of rationality but whether he rationally should do so. Conformity with the independence postulate relative to a value neutral mixture set generated by the menu as a matter of rationality may, indeed, be trivial as a prescription. So is the requirement that a rational agent ought to be certain of the deductive consequences of the set of propositions of which he or she is certain. This form of triviality is a virtue in a prescriptive theory of rationality even though compliance with the prescriptive theory is far from trivial.

Sen is right to observe that expected utility principles are unsuccessful as predictive and descriptive models of human behavior in many circumstances. People lapse from standards of rationality. Moreover, as a prescriptive principle, the injunction to maximize expected utility is applicable only under conditions that depend on the contextual factors that are not always explicitly specified in the formula for expected utility. Sen acknowledges that expected utility theory is, when compared to alternatives, relatively successful (176, fn. 29). He seeks to identify its limitations. I agree that it has limitations both as descriptive theory and prescriptive principle. Chang's evaluation of the option of tossing the coin to determine which child is treated ought not depend exclusively on the expected utility calculation for coin tossing.[9]

[9] It may be worth mentioning that if Chang had at his disposal an alternative to a stochastic process for making the decision such as the advice of a seer or the dictates of the law, he might avoid responsibility for what he does in that way. These alternatives would also be rationally coherent even if morally questionable.

But this limitation undermines neither the expected utility principle nor the independence postulate.

In any case, whether we consider "choice consistency conditions" or "independence" postulates or other candidate requirements on rational choice, it is clear that requirements of rationality depend for their legitimacy not only on the menu of options available but on the decision maker's value commitments and beliefs. I have quibbled here with some of the respects in which Sen has brought this point to bear on various types of applications. But his diverse elaborations are unfailingly instructive and supportive of his general thesis that rationality is a discipline and neither a favored formula, nor an essentialist doctrine.

II. SELF-INTEREST MAXIMIZATION, THE PRISONER'S DILEMMA, AND THE ASSURANCE GAME

Sen has contended for a long time that the prisoner's dilemma and its more sophisticated variants can be understood as challenging the idea that rational agents should pursue their own goals "subject to feasibility considerations, without being restrained by any other values" (216). Sen makes clear that the goals challenged are not merely those restricted to promoting individual consumption without sympathy for others. Nor are they limited to efforts aimed at promoting one's own welfare where account is taken of the benefit accruing from sympathy with the benefits accruing to others (or the losses resulting from sympathy with the misery of others). The challenge Sen means to raise is against *self-goal choice*. "Each act of choice of a person is guided immediately by the pursuit of one's goal (and in particular, it is not restrained by the recognition of other people's pursuit of their goals)" (214).

Departing from self-goal choice is not to be confused with changing one's goals by incorporating into them consideration of the goals of others. To do the latter is changing one's goal. But the new goal, like the old one, is the agent's own goal. To be guided by that goal continues to be acting according to self-goal choice. Rejecting self-goal choice according to Sen appears to amount to this: even though the decision maker retains a given system of goals, the decision maker does not act in a manner that promotes those goals but rather in a fashion that "is constrained by the goals of others, or by rules of conduct, thereby violating self-goal choice (that is, the influences may affect the person's choice without their taking the form of goals that the person can be seen as pursuing himself)" (214).

In order to understand what Sen has in mind, recall the prisoner's dilemma. It is common wisdom that in the one shot prisoner's dilemma, rational players both defect because defecting is better than cooperating no matter what the other player does. Sen shares this common wisdom. Common wisdom is sensible if both players judge the other player's choice to be probabilistically independent of what he chooses. But player 1 may think it highly likely that player 2 will choose like he does. In that case, player 1 could find it better to cooperate than to defect. If player 2 did have similar kinds of beliefs to player 2, both players could perfectly rationally pursue their own goals and cooperate.

Sen is aware of this view but he does not seem to take it seriously himself. He thinks defection is rationally mandatory in the prisoner's dilemma provided that each player acts in a manner that is maximal given the agent's goals.

Sen seems to think that the players pay a price for self-goal choice. Both players acknowledge that they would be better off if they both cooperated.

I do not see that the players pay any price. Neither player has control over what the other player does. Each player's options are restricted to what he does. So even though both players recognize that both cooperating is pareto superior to both defecting, this fact is irrelevant since no agent has control over whether they jointly cooperate or jointly defect.

Sen thinks the lack of control can somehow be finessed. Sen reasons that "if the recognition that we can all better pursue our respective goals by jointly departing from goal priority [self-goal choice] makes us do exactly that, why should that departure change the nature of the goals that we are trying to pursue?" (212). Sen suggests that this sort of result can be brought off if both parties to a prisoner's dilemma acted *as if* the payoff structure of the game were different. In the payoff structure of the prisoner's dilemma, each player prefers his defecting and the other player cooperating to both cooperating and prefers both defecting to his cooperating and the other defecting. Sen suggests acting as if the payoff structure were that of the Assurance Game. Each player prefers both cooperating to his (or her) defecting and the other cooperating and prefers both defecting to his (or her) cooperating while the other defects. There are two equilibrium solutions to the Assurance Game game (both cooperating and both defecting).

The goals characterized by the payoffs in the Assurance Game do not represent the goals of either player. But suppose they act as if these were their goals. Moreover, suppose that the two players have enough confidence in each other to judge it likely that if their partner will cooperate, then they will both cooperate. Sen argues that both players would be better off according to their true goals acting as if their goals were those spelled out for the Assurance Game than they would be if they both acted by promoting the true goals. So they should not act in accordance with their true goals (those represented by the payoff structure of the prisoner's dilemma) but in accordance with the payoff structure of the Assurance Game.

Two objections come immediately to mind. The first is that the argument for acting in accordance with the payoff structure of the Assurance Game depends on each player having confidence that the other will cooperate—that is, on a prediction as to what the other player will do that cannot be derived solely from considerations of the other player's rationality and information about the predicament. Trust or confidence in the other is required as well. But as noted above, using the payoff structure of the prisoner's dilemma leads to cooperation if each player believes that the other player will choose in the way he does. So even if pretending that one's goals are different than they are were reasonable, it is not necessary once the centrality of confidence or trust in the other is appreciated.

Even if this point is waived, however, it is difficult to understand why the use of the payoff structure of the Assurance Game does not represent a change in each player's goals rather than an abandonment of self-goal choice. Perhaps prior to making the change, each player might reason that in the light of his current values and goals (for example, those represented by the payoff structure of the prisoner's dilemma) it would be best to alter his goals. This argument does not justify abandoning self-goal choice. In general, whether one is seeking to justify a change in beliefs or in values, the justification should be grounded in the beliefs and values prior to change. But once one implements the change one has concluded is to be made, the beliefs and values one endorses are different from what they were. But they are not "as if" beliefs and values.

In sum, I agree with Sen that rational agents can cooperate in prisoner's dilemmas—even in the absence of prior agreements and understandings. This view is in keeping with Sen's ecumenical attitude towards rationality. But I do not understand how a rational agent can do other than choose in a way that promotes the goals and values the agent endorses in the context of deliberation. Of course, rationality does not require that the goals and values be restricted to "self-centered welfare" or to what promotes a self-welfare goal maximizing the welfare of the agent even if the agent is benevolent and includes the interests of others in his or her welfare. It seems that Sen's tendency to think that we need to break not only with self-centered welfare or self-welfare but with self-goal choice derives from the failure to appreciate the possibility of changing one's goals and values for a reason while continuing to satisfy the requirements for self-goal choice.

A rational agent should choose an option that is optimal according to the agent's point of view if that point of view recognizes the availability of an optimal option. Other points of view do not count unless the agent makes those other points of view his own.

III. INCOMPLETE PREFERENCE

Deliberate (as opposed to routine or automatic) decision making requires that the decision maker ascertain enough information about his or her value commitments and beliefs as they relate to the menu of options S so as to identify a value structure $V(S)$ for S. Condition (ii) requires that the value structure $V(S)$ for menu S must be a weak order that guarantees that weak preference is reflexive, transitive, and complete, equipreference is reflexive, transitive, and symmetrical, strict preference is irreflexive, transitive, asymmetrical, and that strict preference and equipreference are connected: for every x and y in S xPy, xIy, or yPx.

Sen rightly challenges the thesis that rationality requires that $V(S)$ (or, indeed, $V(\Omega)$) should be a weak ordering.

> A chooser, who may have to balance conflicting considerations to arrive at a reflected judgment, may not, in many cases, be able to converge on a complete ordering when the point of decision comes. If there is no escape from choosing, a choice decision will have to be made even with incompleteness in ranking (160).[10]

Sen has been one of the rare economists (or philosophers for that matter) who has understood, rightly in my view, that "incompleteness in the ranking" may not be eliminated by the moment of choice.[11] The decision maker X may have to face a choice where, given the best information available to X at the time of choice, there is no optimal option—no option in the set of available alternatives at least as good as every other alternative. Sen correctly maintains that being confronted with such a predicament is not a mark of irrationality. To

[10] In a footnote, Sen writes that incompleteness can arise from "limited information" or "unresolved conflict."

[11] Indeed, incompleteness in preference or evaluation is not widely studied by psychologists and social scientists interested in decision making. As a consequence, experimental data are misleadingly interpreted. For example, it is conventional wisdom that experimental data has conclusively established widespread violation of the independence postulate or sure thing principle. That is true only if the behavior observed is taken to reveal definite preferences rather than choice in the face of unresolved conflict.

the contrary, it is a mark of honesty and good sense to refuse to "make up one's mind" when reflection on "all things considered" has not provided a warrant for doing so.

In this connection, Sen draws an important distinction between "tentative incompleteness" and "assertive incompleteness." The decision maker faces a decision problem with tentative incompleteness in his value structure if the agent lacks the opportunity to pursue matters further and must make a decision even though he or she thinks that given sufficient time, opportunity, and resources, a warranted resolution removing the incompleteness would be forthcoming. Assertive incompleteness obtains if the decision maker despairs of the possibility of resolving the conflicts in his or her value commitments even "in principle." Sen wishes to provide room for the sort of tragic choices that Bernard Williams, Martha Nussbaum, and Charles Larmore and other acolytes of Isaiah Berlin insist upon. I agree that principles of rationality ought not to preclude such predicaments. But as a matter of methodological morals, it seems to me that we should avoid placing roadblocks in the path of inquiry as Peirce would say. No unsettled issue either in science, morals, or politics should be taken to be irresolvable in principle unless we have an impossibility theorem. Otherwise we should continue to entertain the "hope" as Peirce would say that future inquiry could bring resolution for similar kinds of issues.[12]

Sen's distinction and the issues it raises are philosophically important. They go to the heart of the question as to the extent to which problems in ethics are amenable to inquiry. But as far as the question of rational choice is concerned, the topic may be bypassed as Sen clearly recognizes. Whether incompleteness at the moment of choice

[12] See Levi, "Conflict and Inquiry," *Ethics*, CII (1992): 814–34, reprinted in *The Covenant of Reason* (New York: Cambridge, 1997), chapter 11. It is important not to confuse Sen's distinction between tentative incompleteness and assertive incompleteness with another important contrast between imprecision and indeterminacy. Agent *X* may be committed to a value structure $V(S)$ representable by a weak ordering over *S*. Yet *X* may not be able to identify fully that weak ordering with complete precision. *X* has failed to fulfill *X*'s commitments. There is incompleteness in *X*'s "performance." At the moment, *X* cannot explicitly indicate whether *X* prefers *a* over *b*, *b* over *a*, or equiprefers *a* and *b*. Yet, there is neither tentative nor assertive incompleteness in *X*'s value structure. By way of contrast, *Y* might be able to identify *Y*'s value structure as one that is incomplete. We may say that *X* is ignorant of the details of *X*'s value structure. *X* is in need of therapy, training, or aid in computation to fulfill *X*'s commitment. *Y* is fully aware of the incompleteness in *Y*'s value structure. There is no need for therapy, training, and so forth. *Inquiry* will, if successful, remove tentative incompleteness. Neither therapy nor inquiry can remove assertive incompleteness. Needless to say, agent *Z* might be committed to an incomplete value structure and not be able to identify it precisely.

is tentative and the conflict in values can in principle be resolved or is assertive in a way that precludes resolution, the exigencies of choice require deciding without resolving the conflict. In order to provide a framework for addressing such predicaments, the account of value structure for the menu of options given in condition (ii) needs to be amended. The requirement of completeness must be removed from the condition of a weak ordering.

What, however, is entailed by giving up completeness? Recall that for Sen and many others, a complete ordering is one representable by a weak preference relation *R* that is reflexive, transitive, and complete and where strict preference is the asymmetric factor of *R* in the sense that *xPy* if and only if *xRy*&~*yRx* while equipreference is the symmetric factor so that *xIy* if and only if *Xry*&*yRx*. Sen calls the weak preference with reflexivity and transitive but without completeness a *quasi ordering*. The factorization of weak preference into strict preference and equipreference is taken to be a matter of definition so that when completeness of weak preference is given up, Sen retains the definitions of strict preference and indifference. However, Sen's practice here is very much open to question for the same reasons that justify calling the requirement of completeness into question.

Incompleteness in preference arises when the value commitments, goals, and values of the decision maker impose constraints on the way that the decision maker ranks the options available in *S* that allows several different rankings to be permissible. *X* has promised to pay *Y* a certain sum of money by a certain date in return for specified services. *Y* renders the services and *X* confronts the decision whether to pay the sum in cash or by check or to refuse to pay the bill. *X* acknowledges the obligation to keep promises and, hence, ranks refusal to pay below the other alternatives. How should *X* rank the other alternatives? An obligation to keep promises may not impose any constraint on how *X* ranks these options. Three rankings are permissible:

Way 1: Paying by check and by cash are equipreferred (and both preferred to refusal to pay).

Way 2: Payment by check might be ranked above payment in cash and paying by cash over nonpayment.

Way 3: Payment in cash may be ranked over payment by check that is again ranked over nonpayment.

The obligation to keep promises and pay debts rules out none of these rankings. If *X*'s value commitments impose no additional constraints, *X* is committed to recognize all three rankings as permissible rankings of these options. Of course, *X* could endorse further value commitments that rule out one or more of these permissible rankings as impermissible. Ways 2 and 3 might be ruled out because *X* judges it a matter of indifference whether he pays by check or by cash. *X* might wish to help *Y* evade taxes and rule out ways 1 and 2. *X* might simply think that it is not his obligation to impede *Y*'s tax evasion so that *X* rules 2 but neither 1 nor 3. *X* could see *X*'s duty to be to impede such tax evasion and rule out ways 1 and 3. Or *X* might think that it is not his obligation to impede such tax evasion and thus rule out way 3 alone.

X cannot, however, reasonably rule out 1 while retaining 2 and 3 as permissible. If *X* is in doubt as to whether paying by cash is better than paying by check or the other way around, then ranking the two options together should be permissible as well. The latter ranking is a "potential compromise" between the other two and should be regarded as permissible if they are.

In any case, if all three ways of evaluating the options are recognized to be permissible, the only constraint *X* endorses on *X*'s evaluations is the obligation to keep *X*'s promises. *X* clearly recognizes that both paying in cash and by check are strictly better than refusing to pay. Every permissible ranking agrees on that point. But paying by cash and paying by check are noncomparable. Neither option is weakly preferred to the other. And no option is weakly preferred to every other option. According to standard definitions of optimality including Sen's, there is no optimal option among the three available. *X* cannot optimize even though *X* is rational.

Now the notion of preference on which I relied to describe this illustration characterizes a value structure in terms of the weak orderings or rankings of the options recognized as permissible according to the constraints generated by *X*'s value commitments. Option *x* is categorically weakly preferred to option *y* if and only if *x* is weakly preferred to *y* according to all permissible rankings in the value structure. *x* is categorically strictly preferred to *y* if and only if *x* is strictly preferred to *y* according to every permissible ranking and is categorically equipreferred to *y* if and only if *x* is equipreferred to *y* according to every permissible ranking.

The notion of categorical weak preference deployed here is formally similar to Sen's notion of a binary relation of aggregation R^a in chapter 7* of *Collective Choice and Social Welfare* and the notion of a weak preference derived by "utility based intersection" discussed on pp. 362–63 of chapter 11 of the book under review. In developing a basis for classifying different kinds of interpersonal comparisons, Sen considers different rankings of the social states or alternatives derived from different ways of "aggregating" the preferences or valuations of the individual citizens or beneficiaries of social welfare. The details do not matter here. It is sufficient to notice that a given family of such rankings in the setting of social choice corresponds to a family of permissible rankings in the value structure of an individual decision maker. The weak preference relation R^a is defined in a fashion entirely analogous to the way I defined categorical weak preference.

There is, however, one important difference between how Sen thinks of incompleteness in valuing or preference and incompleteness in categorical preference. According to Sen, strict preference and equipreference are obtained by factoring weak preference into its asymmetric and symmetric parts. Categorical preference as I have defined it is not always factorizable in this fashion.

Return to our example. Suppose *X* rules way 3 out as impermissible. The other two ways are recognized as permissible. In that case, paying by check is categorically weakly preferred to paying by cash since it is weakly preferred according to every permissible ranking. However, it is neither categorically equipreferred nor categorically strictly preferred to paying by cash.

However, it is also true that paying by cash is not categorically weakly preferred to paying by check. According to Sen's approach, paying by check is strictly preferred to paying by cash. Whereas I derive the claim that *X* does not categorically strictly prefer paying by check over paying by cash from the given value structure and categorical weak preference, Sen derives exactly the opposite from the *same* value structure and categorical weak preference. In effect, Sen's approach reduces the value structure that recognizes ways 1 and 2 as permissible and 3 as impermissible to a value structure where only 2 is recognized to be permissible. Sen does this by resorting to the standard practice of defining strict preference as the asymmetric component of weak preference.

According to both approaches, the categorical weak preference is a quasi ordering. But Sen's approach sees the quasi ordering as free of any indeterminacy. Strict preference and equipreference are connected. Given any pair of options in the triple, either one is strictly preferred to the other or they are equipreferred. In facing a three way choice, Sen's view recognizes paying by check as uniquely optimal and maximal. An option is maximal if there is no option in the menu of available options strictly preferred to it. This is Sen's favored criterion for choice when there is incompleteness in the value structure. Paying by check is strictly preferred to paying by cash according to Sen's understanding of strict preference. According to Sen's recommended approach, paying by cash is not admissible for choice.

According to the approach I use, paying by check is uniquely optimal but it is not uniquely maximal. Paying by check is not categorically strictly preferred to paying by cash. Maximality, Sen's recommended criterion for choice, allows for the choice of either option.

Thus, the difference here does not concern how one should make choices in the face of incompleteness in one's valuations or preferences. I am assuming Sen's principle of maximality. The problem arises because Sen understands the relation of aggregation or categorical preference as taking strict preference to be the asymmetric factor of weak preference. The difference between Sen's approach and the one I favor is no issue of terminological legerdemain. Sen's practice rules out the relevance of equipreference between a pair of options *a* and *b* as a permissible ranking when there is a permissible strict preference for *a* over *b* but not the other way around. Sen is prepared to recognize certain kinds of incompleteness in value structure. And his championing this view and appreciation of its significance is a first rate contribution. My reservation with his account concerns one of those rare cases where Sen has failed to be ecumenical enough in recognizing as rational certain kinds of value commitments.

Given Sen's own ambition to take a generous, comprehensive view of rationality, it seems to me that when he rightly acknowledges the rationality of having incomplete preferences, he should also recognize the rationality of sometimes having complete but disconnected preferences as well. As matters stand, Sen denies the rational coherence of the latter idea.

IV. MAXIMALITY VERSUS V-ADMISSIBILITY

The value structures used to derive the weak preference quasi orderings to which maximality applies can themselves be derived from sets of expected utility functions in ways that need not detain us here but that provide a way of accommodating decision making under uncertainty within the purview of the current discussion. In this setting, maximality as a criterion of rational choice has the virtue of, on the one hand, allowing for incomplete preferences and complete but disconnected categorical preferences while, on the other hand, prohibiting, in agreement with the classical expected utility theory, choosing options that are "dominated" by other options.

But maximality is not the only candidate criterion of rational choice that exhibits this double-barreled virtue.

Call an option *V*-admissible in a menu set *S* if and only if there is a permissible utility function in the value structure *V*(*S*) according to which the option is ranked as optimal. All *V*-admissible options must be maximal. There can be no option in *S* that is categorically strictly preferred in S to an option that is V-admissible in S. But there can be maximal options in S that are not V-admissible. An abstract example will have to suffice here. Imagine a choice between three options x, y, and z. According to one permissible utility function in $V(S)$, $v(x) = 1$, $v(y) = 0.4$, and $v(z) = 0$. According to another, $v(z) = 1$, $v(y) = 0.4$, and $v(x) = 0$. The set of permissible utility functions is the set of all weighted averages of these two (and positive affine transformations of these). y cannot come out optimal according to any such utility function. Only x and z can. So x and z are V-admissible but y is not. Yet y, along with x and z are maximal.

So we have two candidate criteria for choice that allow incompleteness in ordering options in S and respect the independence postulate. Is there any basis for deciding between them?

In the ecumenical spirit that is exhibited by Sen's attitude towards rationality, one could argue in favor of maximality on the grounds that it allows more options than V-admissibility does to be admissible given the value commitments and beliefs of the decision maker.

But the *same* ecumenical spirit should allow us to recognize an important difference between the value structure for the example just given and a variant of it where $v(y)$ is given 0.6 in both rankings. We should allow rational agents to be sensitive to such differences between second worst and second best cases. Maximality does not do well by this standard. It *forbids* taking such differences to be relevant in decision making. V-admissibility, on the other hand, has resources sufficient to distinguish between agents whose value commitments do not recognize a relevant distinct between second worst and second best cases and agents whose value commitments entail such recognition. V-admissibility countenances the rationality of both kinds of agents.

V. CHOICE CONSISTENCY AGAIN

Throughout this discussion, I have been assuming that whether or not value structures for a universal set Ω and nonempty finite subsets of Ω exhibit preference independence and that whether the choice functions are derived using maximality or V-admissibility, they will be menu independent as well.

Under this assumption, maximality, as Sen acknowledges, will violate β. It will, however, conform to α and γ. Given a class M of option or menu sets, x is in the maximal set for every S in M if and only if it is in the maximal set for the union of all sets in M. When preference is menu independent, Sen is willing to acknowledge that rational agents ought to conform to α and γ.

Advocates of V-admissibility are in this respect harsher critics of choice consistency requirements than Sen appears to be. Even when preference is menu independent, choice functions defined by *V*-admissibility violate γ.[13] This is another respect in which *V*-admissibility is more congenial with a comprehensive, ecumenical approach to rationality than maximality.

VII. CONCLUSION

As I have already indicated, Sen's book focuses as explicitly as many of Sen's books do on devising ways and means of adapting the analytic devices favored by economists to furnish a way to incorporate a serious recognition of moral and political values within the framework of economic analysis—especially social choice theory. As I indicated at the outset, I have not touched at all on the rich treatment of how considerations of freedom may be accommodated within the framework of social choice theory. And only the surface of Sen's approach to developing a comprehensive ecumenical approach to rationality has been scratched.

I have argued that although Sen's attack on excessively formulaic requirements on rationality such as internal choice consistency is welcome and the idea of menu dependent preference is an important topic for critical scrutiny, more motivated accounts of choice consistency based on maximizing in Sen's sense in a setting of menu independent preference insist on requirements for choice consistency that are stronger than what rationality should require.

The cogency of this complaint depends in large measure on the contention that maximality is not comprehensive and ecumenical enough as a principle of rational choice when preferences are incomplete.

Having said this, it must be emphasized that the dispute between advocates of maximality such as Hans G. Herzberger, Sen, and Peter Walley[14] and those sympathetic with *V*-admissibility is a controversy among allies sharing a common preoccupation with forging a framework for normative principles of rationality tolerant of diversity while insistent on a respect for logicality and reason. Sen's book is testimony to his abiding, serious commitment to and insightful and rigorous pursuit of this vision.

ISAAC LEVI

Columbia University

[13] Levi, *Hard Choices*.

[14] In addition to Sen, Herzberger, "Ordinal Preference and Rational Choice," *Econometrica*, XLI, 2 (March 1973): 198–99; and Walley, *Statistical Reasoning with Imprecise Probabilities* (London: Chapman and Hall, 1991), pp. 162–66.