(to appear in Philosophical Studies)

Physicalism from a Probabilistic Point of View

Elliott Sober*
Philosophy Department
University of Wisconsin, Madison 53706

Physicalism -- like other isms -- has meant different things to different people. The main physicalistic thesis I will discuss here is the claim that all occurrences supervene on physical occurrences -- that the physical events and states of affairs at a time determine everything that happens at that time. This synchronic claim has been discussed most in the context of the mind/body problem, but it is fruitful to consider as well how the supervenience thesis applies to what might be termed the organism/body problem. How are the biological properties of a system at a time related to its physical properties at that time?

Philosophers have discussed the meaning of supervenience claims at considerable length, mainly with an eye to getting clear on modal issues. Less ink has been spilled on the question of why anyone should believe this physicalistic thesis. In what follows, I'll discuss both the metaphysics and the epistemology of supervenience from a probabilistic point of view. The first half of this paper will explore how supervenience claims are related to other issues; these will include the thesis that physics is causally complete, the claim that there are emergent properties, the idea that mental properties are causally efficacious, and the notion that there are scientific laws about supervenient properties that generalize over systems that deploy different physical realizations of the properties in question.. The second half will examine the question of how observational evidence can lend support to supervenience claims. This problem turns out to raise some surprisingly deep issues about the nature of hypothesis testing in science.[1]

1. Preliminaries

What, exactly, is the supervenience thesis supposed to mean? It adverts to "physical" properties, but what are these? This question apparently leads to a dilemma (Hempel 1969, Hellman 1985, Crane and Mellor 1990). If the supervenience thesis presupposes that the list of fundamental magnitudes countenanced by present physics is true and complete, then the claim is almost certainly false, since physics is probably not over as a subject. Alternatively, if we interpret "physical property" to mean what an ideally complete physics would describe, then the supervenience thesis lapses into vagueness. What does it mean to say of some future scientific theory that it is part of the discipline we now call "physics?"

Papineau (1990, 1991) suggests one way to address this dilemma. Rather than worry about what "physics" means, we should instead exploit the fact that we know what it is for a

property to be mental. Human beings have beliefs, desires, and sensations, but the very small physical parts (cells, molecules, atoms, etc.) of which human bodies and their environments are made do not. The claim of interest is that mental properties supervene on non-mental properties. In reply, it should be noted that worries about the meaning of "physical property" also can be conjured up in connection with the concept of the mental. Our current list of mental properties -- both those used in folk psychology and the ones at work in science -- may well undergo change. Can we really say with any precision what makes a future scientific predicate "psychological?"

I think the right response is simply to admit that we know how to apply the concepts of mental and physical property only in so far as the properties in question resemble those that we currently call mental and physical. If future science departs dramatically from current conceptions, it may be impossible to say whether that science ends up vindicating or overturning what we now call physicalism. However, this possible indeterminacy does not deprive physicalism of its philosophical interest. There is still a point to the question of whether psychological properties, as we currently understand them, supervene on physical properties, where the latter are understood in terms of our current best physical theories. Current physics does not include belief, desire, or sensation in its list of fundamental properties. Do these properties supervene on space-time curvature, mass-energy, and the other properties that are discussed in current physics? We miss an interesting philosophical question if we reject the question by pointing out that current physics is probably wrong or incomplete in some respect we know not what.

Above, I described supervenience as endorsing a determination claim. A system's mental properties at a time are a function (in the mathematical sense of a mapping) of its physical properties at that time. This way of putting the point leaves open that the converse relation may also obtain -- that a system's physical properties are a function of its mental properties. This raises the question of whether the concept of supervenience offers any prospect of identifying an asymmetry between the mental and the physical.

It is tempting to use the idea of multiple realizability to argue that a system's mental properties don't determine its physical properties. If a system can have mental property M by having any of several physical properties P_1, P_2,..., P_n, then each P_i determines that M will be present, but M does not determine which of the P_i is present. Quite so, but this settles the matter only if one assumes that disjunctions of physical properties aren't physical properties; after all, M does guarantee that the disjunction (P_1 or P_2 or ... or P_n) is instantiated.

A simpler and less metaphysically weighty way of locating an asymmetry is available. Even if multiple realizability were false, the physical still would not, in general, supervene on the mental. The reason is that there are lots of physical things that don't have any mental properties at all. The sun and Lake Michigan are alike psychologically -- neither has beliefs, neither has desires, neither has sensations, and so on. Yet, they differ physically. Even if there were a one-to-one mapping between mental and physical properties in the domain of individuals who have minds, there is no such mapping that covers mindless individuals. From this more general

perspective, in which both minded and mindless individuals are taken into account, supervenience involves an asymmetry. A system's physical properties uniquely determine its mental properties, but not, in general, conversely.[2]

I have described supervenience as a synchronic relation between instantaneous states -- the physical properties at time t determine the nonphysical properties at time t. However, a moment's reflection shows that many of the properties we attribute to systems are temporally "thick." For example, when we say that Sally saw the moon at time t, we are describing a relationship between what is going on in Sally's mind at time t and the state of the moon some time earlier. Her seeing the moon entails that a process took place that lasted from t-dt to t. A supervenience thesis should cover attributions of this sort. The easy solution is to let talk of "time t" denote intervals as well as instants. Supervenience remains a synchronic relation, and so it contrasts with the similar sounding diachronic relation of causal determination.

What I've just said about time also applies to space. Philosophical discussion of supervenience has examined the question of how much of the physical world at a time is needed to determine the state of a given supervening property that attaches to an individual at the same time. Surface grammar is a poor guide to this issue. Biologists talk about the fitness of organisms, and philosophers of biology have often claimed that fitness supervenes on physical properties. However, the supervenience base has to include properties of the environment as well as properties of the organism. The fitness of an organism supervenes on the properties of a larger, containing, system. Parallel points have been emphasized by philosophers of mind who think that the semantic content of a mental representation is "wide" -- that it depends on features of the organism's environment. My discussion in what follows won't be affected by how large the containing system has to be when one considers whether this or that property of an organism supervenes on properties of the containing system. For convenience, I'll usually talk of a system's mental or biological properties supervening on the system's physical properties, but this is solely for the sake of convenience.

2. Supervenience entails the causal completeness of physics, but not conversely.

Let M = a mental property of a system at time t, P = all the physical properties of the system at time t, and B = a behavior of the system at time t+dt. I understand the supervenience claim to assert that:

(S) $\Pr(M \mid P) = 1.0$.

That P confers on M a probability of 1.0 is a reasonable representation of the idea that P necessitates M, as long as there are finitely many states that M and P might occupy; if P is true, there is no chance that M will be false.[3]

The thesis that physics is causally complete says that

(CCP) $Pr(B \mid P) = Pr(B \mid P\&M)$.

That is, the chance at time t that B will occur at time t+dt is fixed by the physical properties that the system has at time t; the value is unaffected by taking account of the system's mental properties at time t as well.[4] (CCP) says that the physical properties instantiated at time t "screen off" the mental properties instantiated at that time from behaviors that occur afterwards. Although (CCP) is an ontological, not an epistemological, thesis, it has a natural epistemological reading -- knowing the mental properties of the system is superfluous, once you know the physical properties, if the goal is to predict whether the behavior will occur.

Thus defined, supervenience entails the causal completeness of physics. To see why, consider the following expansion of the expression $Pr(B \mid P)$:

$$Pr(B \mid P) = Pr(B\&P) / Pr(P)$$

$$= [Pr(B \& P \& M) + Pr(B \& P \& \text{not-}M)] / Pr(P)$$

$$= [Pr(B \mid P\&M)Pr(P\&M) + Pr(B \& \text{not-}M \mid P)Pr(P)] / Pr(P)$$

$$= Pr(B \mid P\&M)Pr(M \mid P) + Pr(B \& \text{not-}M \mid P)$$

From this last equation, it is clear that if $Pr(M \mid P) = 1.0$, then $Pr(B \mid P) = Pr(B \mid P\&M)$.

What about the converse? Does (CCP) entail (S)? The answer is no. To see why, suppose that supervenience is false, in that $0 < Pr(M \mid P) < 1.0$. This allows us to expand $Pr(B \mid P)$ as follows:

$$Pr(B \mid P) = Pr(B \mid P \& M)Pr(M \mid P) + Pr(B \mid P \& \text{not-}M)Pr(\text{not-}M \mid P).$$

If we substitute r for $Pr(B \mid P)$, for $Pr(B \mid P\&M)$, and for $Pr(B \mid P \& \text{not-}M)$, as (CCP) allows, we obtain

$$r = r\,[Pr(M \mid P)] + r\,[Pr(\text{not-}M \mid P)].$$

Notice that this equation is true, regardless of the value assigned to $Pr(M \mid P)$.

3. The physical detectability of the mental

The principle of the physical detectability of the mental says that if the mental property M (or its negation) is instantiated by an individual at time t, then a suitably situated physical device would be able to detect the presence or absence of M at t by registering a physical property D or its negation at some later time t+dt. I suggest that detectability be given a probabilistic reading:

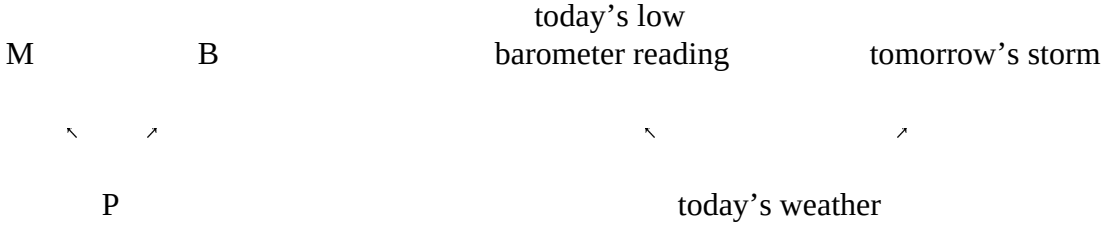
(PDM) $\Pr(D \mid M) \neq \Pr(D \mid \text{not-}M)$.

When the probabilities in (PDM) take on the extreme values of 1 and 0, it is possible to establish with certainty whether M was exemplified at t by observing whether D is exemplified at time t+dt. When the probabilities are more intermediate, the inference from later physical state to earlier mental state becomes more chancy.

Papineau (1990, 1993, 1995) and Loewer (1995) argue that (PDM) and (CCP) together entail (S). If B at time t+dt is a detector of M at time t, and P fixes the probability at time t that B will be true at t+dt, then M must supervene on P, or so they argue. In fact, the implication does not go through. To see why, it is useful to think of the relations that connect P, M, and B on the model of Reichenbach's (1956) principle of the common cause. Reichenbach specified the probability relations he thought must obtain when two events are correlated because they trace back to a common cause. It doesn't matter whether Reichenbach was right in everything he said about causality; it also doesn't matter that P is not a common cause of M and B (recall that P and M are simultaneous events). The point is that Reichenbach described a set of formal probability relationships that may be applied to the case at hand.

Reichenbach proved that if P raises the probability of M and does the same for B, and if P renders M and B conditionally independent of each other (i.e., screens off each from the other), then M and B will be positively correlated. There is no need for states of P to confer probabilities of 1 and 0 on states of M, if P is to screen off M from B and M and B are to be correlated. (CCP) and (PDM) can be true even when (S) is false.

The following figure illustrates the parallelism I am exploiting between a common cause set-up and the relationship of P, M, and B:



Just as (CCP) says that P screens off M from B, so it is true that Pr(storm tomorrow | today's weather) = Pr(storm tomorrow | today's weather & today's low barometer reading). Just as (PDM) say that M and B are correlated, so it is true that Pr(storm tomorrow | today's barometer reading is low) > Pr(storm tomorrow | today's barometer reading is not low). Yet, it does not follow from all this that Pr(today's low barometer reading | today's weather) = 1.0. (S) does not follow from (CCP) and (PDM).[5]

(PDM) can be deduced from (S) and (CCP), once these latter propositions are strengthened. (S) says that a system must exhibit M if it has P. But what will happen if the system lacks P? Let us add the claim that P makes a difference in whether M is true:

$$Pr(M \mid P) \neq Pr(M \mid \text{not-}P).$$

Similarly, (CCP) says that M is irrelevant to predicting B, once you know that P is true; (CCP) neglects to say that P is actually relevant to predicting B. So let us add that further claim:

(CRP) $Pr(B \mid P) \neq Pr(B \mid \text{not-}P)$.

With these additional premisses, we can deduce (PDM). The derivation is just the one that Reichenbach described.

Although (CCP) and (S) are general principles that are part of what physicalism asserts, (CRP) is not. There may well be events that occur at random, relative to all the physical facts that were true at an earlier time. Physicalism, as I understand it, is not committed to denying that this is possible.

4. Emergent causation and diachronic determinism

Let the thesis of emergent causation assert that the occurrence of the mental property M at time t makes a difference in the probability that behavior B will occur at time t + dt, even after all the physical properties P at time t are taken into account:

(EC) $Pr(B \mid P\&M) \neq Pr(B \mid P)$.

(EC) is the denial of (CCP). It follows that (EC) must be incompatible with any proposition that entails (CCP) -- specifically with (S). Causal emergentists must deny supervenience.

How is the synchronic thesis of supervenience related to the diachronic thesis of causal determinism? We have already seen that if $Pr(M \mid P) = 1$, then $Pr(B \mid P) = Pr(B \mid P\&M)$. That is, supervenience entails that the physical facts at time t "screen off" the mental facts at time t from behaviors that occur at t + dt. However, nothing follows from this as to what value $Pr(B \mid P)$ has; it could be unity, and it could also be less. Supervenience does not entail causal determinism. If so, causal indeterminism does not entail that the supervenience thesis is false (contrary to Crane and Mellor 1990).

Does the conjunction of causal determinism and (CCP) entail supervenience? The answer is, again, no. Even if $Pr(B \mid P) = Pr(B \mid P\&M) = 1.0$, $Pr(M \mid P)$ can still take on any value at all.

5. Different definitions of "emergentism"

What I have called emergent causation -- that mental properties make a causal difference to behavior, above and beyond the difference made by physical properties -- diverges from some of the ideas that McLaughlin (1992) describes in his article on "British Emergentism" (a position that McLaughlin sees in the work of J.S. Mill, Alexander Bain, George Henry Lewes, Samuel Alexander, Lloyd Morgan, and C.D. Broad). For him, emergentism is principally the thesis that the bridge laws that connect lower- to higher-level properties demand explanation, but are in fact inexplicable; they must be accepted as brute facts (with "natural piety," as the emergentists used to say); Horgan (1993) interprets emergentism in the same way. McLaughlin follows Causey (1977) in holding that bridge laws would not require explanation if they took the form of identity statements. However, if they involve (one-way) conditionals, they are said to "demand explanation." Mechanistic reductionism claims that the laws of physics provide the required explanation, while emergentism denies that this is so.

Although McLaughlin understands physicalism to include the thesis that physics can explain all psycho-physical conditionals, I don't see why the view should be understood in this way. A statement that employs both psychological and physical vocabulary cannot be deduced from a body of theory expressed in purely physical language. This fact alone suggests that physicalists should admit that psycho-physical conditionals don't have purely physical explanations. An additional reason for the physicalist to be circumspect here comes from considering how causal conditionals are explained, regardless of the vocabulary in which they are couched. We explain why E_1 causes E_n by describing an intervening process; the explanation for why E_1 causes E_n is that E_1 causes E_2, E_2 causes E_3 , ..., and E_{n-1} causes E_n. Suppose we take two adjacent links in the causal chain just described and ask: Why does E_2 causes E_3? It may be possible to press the same strategy into service again. But how long can this game of question and answer continue? Perhaps it can go on forever. However, the possibility needs to be faced that, at some point, there simply is no explanation for why E_i causes an adjacent E_j; it just does. If this happens, E_i and E_j may both be physical, both may be mental, or they may stand on either side of the mental/physical divide (assuming for the moment that this is a precise distinction). Physicalism, in its zeal to assert the hegemony of physics, should not assert that every causal connection has a physical explanation; some may have no explanation at all. The same point holds if E_1 and E_n are simultaneous events, with E_n supervening on E_1. The question of why this is so may be raised and answered. But sooner or later, it may happen that there is no explanation for why Ej supervenes on E_i ; it just does. Physicalism, should not claim that every fact of supervenience has a physical explanation; some may have no explanation at all.

In what sense does a psycho-physical conditional "demand explanation?" It may strike us that a given conditional has an explanation, and so we may feel obliged to find out what that explanation is. Here it is the inquirer, not the conditional, that is doing the demanding. Do all psycho-physical conditionals "demand explanation?" That is, should we believe that all such conditionals have explanations? I see no reason to think so, nor do I see why physicalism should

be so committed. Physicalism can say that some conditionals may turn out to have the same status that many now associate with identity statements -- as not having explanations at all.

Another issue that mattered a lot to the emergentists, which McLaughlin discusses but doesn't use to define what he takes emergentism to be, concerns the idea of additivity. Suppose some parts are brought together to form a whole -- molecules of hydrogen combine with molecules of oxygen to form water molecules, for example. If the whole has a property that isn't an additive function of the properties of the parts, then the emergentists said that the property of the whole is an "emergent," not a "resultant."

As the term suggests, additivity is easiest to characterize when the properties of the whole are quantitative. The following 2-by-2 table describes what happens when hydrogen is present or absent and when oxygen is present or absent in a system:

		Hydrogen +	Hydrogen -
Oxygen	+	x+a+b+c	x+a
	-	x+b	x

Putting oxygen into a system when there is no hydrogen there changes the system from x to x+a. Putting hydrogen into a system when there is no oxygen there changes the system from x to x+b. What will happen if we insert both oxygen and hydrogen? If c=0, then the effect of putting the two substances together is just the sum of the effects of putting each of them there without the other; we add (a+b) to the baseline value x. This is what additivity means.

A special case of this formula arises when the system is described in terms of the truth or falsity of propositions; propositions are just dichotomous variables that have two states (the two truth values):

		Hydrogen +	Hydrogen -
Oxygen	+	?	A&-B
	-	-A&B	-A&-B

If A&B would be true when oxygen and hydrogen are both present, the system is additive with respect to the variables ±A and ±B.

Although the concept of additivity raises some interesting philosophical questions, it has nothing to do with what I've called emergent causation. Whether the property of the whole at time t is an additive or a nonadditive function of the properties of the parts at time t, the word "function" means that the properties of the parts determine (via the appropriate composition principle) the properties of the whole. If a value for the constant c is specified, the upper-left cell entry is predictable from the entries in the other three cells; it doesn't matter whether c=0.[6] When the entry in the upper-left hand cell is a function of c and the entries in the three other cells, supervenience obtains and the thesis of emergent causation is wrong. Alternatively, if the properties of the parts at time t confer probabilities that are strictly between 0 and 1 on the properties of the whole at time t, then supervenience fails. This will be true, regardless of whether those probabilities are an additive function of the properties of the parts.

6. Causal efficacy and causal completeness

Although emergent causation (EC) and the causal completeness of physics (CCP) are incompatible, the claim that mental properties are causally efficacious and the claim that physical properties are causally efficacious are perfectly compatible. To see why, we must consider how claims of causal efficacy are understood in a probabilistic framework.

The basic idea in probabilistic representations of causality is that a positive causal factor must raise the probability of the effect, when other simultaneously instantiated causal factors are "held fixed." Conjunctions of these other simultaneous factors are called "background contexts." The idea is that

> C is a positive causal factor for E if and only if $Pr(E \mid C \,\&\, X_i) \geq Pr(E \mid \text{not-}C \,\&\, X_i)$ for all background contexts X_i, with strict inequality for at least one X_i.

The causal claim is assumed to be about the individuals in a specific population. So the idea that smoking is a positive causal factor for the production of lung cancer in the population of human beings alive now in the US means that smoking increases the chance of lung cancer for people with at least one constellation of other properties, and doesn't lower it for any one else. If smoking raises for some and lowers for others, smoking is said to be a "mixed" causal factor, not a positive one (Eells 1991). My point here is not to insist that this probabilistic representation of what causality means is exactly right, but to discuss what is involved in talking about a "background context."

A background context is a conjunction of properties. Inhaling asbestos particles might be one of the conjuncts. If smoking is a positive causal factor for lung cancer, then smoking must raise (or at least not lower) the probability of lung cancer among individuals who have the same level of asbestos exposure. This other factor is "held fixed"; against this homogeneous background, one sees whether smoking makes a difference in the chance of lung cancer.

What one doesn't do is hold fixed the chemical composition of cigarette smoke. To simplify discussion, let's pretend that cigarette smoke always consists of three types of particles (X, Y, and Z) in equal proportions. To find out whether smoking cigarettes is a positive causal factor for lung cancer, one doesn't see whether smoking raises the probability of cancer among people who inhale the same amount of XYZ. By hypothesis, it won't, but that doesn't show that smoking doesn't promote lung cancer. Given the composition of cigarettes in the real world, cigarette smoke automatically contains XYZ. This means that neither is part of the background contexts that have to be considered when the causal efficacy of the other is under discussion.

To see whether inhaling cigarette smoke promotes lung cancer, you investigate whether smoking cigarettes increases the risk of lung cancer among individuals who all inhale the same other things. To see whether inhaling XYZ promotes lung cancer, you investigate whether inhaling XYZ increases the risk of lung cancer among individuals who all inhale the same other things. But if you are studying cigarette smoke, XYZ does not count as "another inhalant." And if you are studying XYZ, cigarette smoke does not count as "another inhalant."

This fact about scientific practice stands on its own, but it also is reflected in the probabilistic criterion described before. If the presence or absence of XYZ is to be part of the background contexts against which the impact of smoking cigarettes (C) on some effect variable is tested (or vice versa), then all conditional probabilities of the form Pr(-- | ±XYZ & ±C) must be well defined. This will not be true if Pr(C & not-XYZ) or Pr(not-C& XYZ) have values of zero.

I hope the parallel point about the relevance of mental and physical properties in the causation of behavior is clear. If I walk from my office to the Student Union, it may be asked why I did so. Consider the hypothesis that this behavior was caused by my believing that the Student Union sells coffee and by my wanting to buy a cup of coffee. To see whether this belief/desire pair is a positive causal factor in the production of that type of behavior, one determines whether the belief/desire pair increases the probability of walking to the Union in different background contexts. These background contexts presumably would include the other beliefs and desires that individuals might have. However, suppose that P_1, P_2, ..., P_n are the (physically specified) supervenience bases that provide different realizations of the belief/desire pair. It would be a mistake to judge the causal efficacy of the belief/desire pair by seeing whether it raises the probability of strolls to the Union among people who have a given P_i.

Thus, the idea that physics is causally complete and that the mental supervenes on the physical does not entail that mental properties are causally inert. In fact, (CCP) and (S), when brought in contact with a probabilistic conception of causality, explain why a certain argument against the causal efficacy of mental properties is invalid. One cannot argue that a property M plays no causal role in producing B by showing that P screens off M from B, if in fact P is a supervenience base of M.

These remarks are not intended to explain how the semantic properties of mental states

manage to be causally efficacious. Rather, they bear on how that problem should be formulated. The issue isn't how mental features at time t can make a difference in what happens later, even after all the physical properties at time t are taken into account. They can't -- not if supervenience is right. But that doesn't show that the mental is causally inert. You don't need emergent causation for mental properties to be causally efficacious; supervenience, the causal completeness of physics, and the causal efficacy of the mental are perfectly compatible, contrary to Kim (1989a, 1989b, 1990).

7. A Harder Question

The idea that one should not hold fixed the supervenience bases of M in assessing M's causal role with respect to B is a straightforward consequence of representing causation probabilistically. However, a harder question is raised by asking what one should hold fixed. To illustrate the puzzle I have in mind, let's modify the story about cigarette smoke. Suppose there are two kinds of cigarettes -- some produce smoke that contains X particles, but not Y, while others produce smoke that contains Y particles, but not X. All cigarette smoke contains Z. Suppose further that X and Y are carcinogenic, but that Z is not. In effect, we can think of cigarette smoke as a disjunction -- (X & Z) or (Y & Z). Should one hold fixed the X and Y levels that cigarette smoke contains in assessing whether smoking causes cancer? Notice that smoke does not entail X (nor does it entail Y), and X does not entail smoke (nor does Y). If one does hold fixed the amount of X and Y that is inhaled, the conclusion will be drawn that cigarette smoke is not a cause of cancer. One will say, instead, that it is X and Y that are carcinogenic.

A parallel line of reasoning leads to the conclusion that the thesis of wide content and the idea of multiple realizability together entail that semantic properties are causally inert. Suppose that the belief/desire pair mentioned in the previous section can have two internal neurological realizations I_1 and I_2, and that a person has the belief/desire pair only if some fact about his or her external environment E holds true. Then the belief/desire pair can be thought of as a disjunction -- (I_1 & E) or (I_2 & E). If one holds fixed I_1 and I_2, the conclusion will be drawn that having the belief/desire pair is not a cause of walking to the Student Union. One will say, instead, that it is the neurological states I_1 and I_2 that cause the behavior.

The question we now face is whether parts of multiply realizable supervenience bases should be held fixed. Eells (1991) says that the background contexts one needs to consider in assessing M's causal role with respect to B must include all properties that are instantiated simultaneously with M, that are independent of M, and that are themselves causally relevant to B. This policy entails that the semantic properties of mental states are causally inert, if content is wide and has multiple neural realizations.[7] It also entails that smoking cigarettes does not cause cancer, if smoke contains both carcinogenic and noncarcinogenic constituents, and the former come in different forms. My hunch is that the right policy for understanding background contexts differs from the one that Eells recommends; unfortunately, I don't know how to formulate what that "right" policy is, let alone defend it. For the moment, it suffices to note that the causal

efficacy of content is threatened no more and no less by this problem than is the carcinogenic character of smoking cigarettes (Segal and Sober 1991).

8. The smart Martian problem

The ideas developed in the previous two sections have an interesting epistemological analog. If the level of X and Y particles inhaled is not part of the background contexts that help define what it means for smoking to be a positive causal factor in the production of lung cancer, then it is a mistake to think that "smoking causes lung cancer" and "inhaling X and Y particles causes lung cancer" are rival hypotheses. A puzzle posed by Robert Nozick (described in Dennett 1987) has its solution in seeing the consequences of this point.

Nozick's smart Martian problem begins with the widely held idea that the justification for attributing beliefs and desires to an individual derives from the fact that such attributions are predictively and explanatorily useful. If we can explain and predict someone's behavior on the basis of mentalistic attributions, and cannot do so otherwise, what better reason could we have for thinking that the attributions are true or approximately so? Conversely, the reason it makes no sense to attribute beliefs and desires to rocks is that these attributions are not needed to explain what rocks do. Mentalistic ascriptions thus seem to stand or fall according to whether they are good inferences to the best explanation. The operative principle seems to be this:

(N) S's behavior provides a reason for attributing beliefs and desires to S if and only if these attributions are needed to predict and explain S's behavior.

Nozick asks us to imagine a race of super-intelligent Martians who can perceive at a glance the states of the elementary particles that make up the bodies of human beings and are able to compute from this information the future states that our bodies will occupy. Since these Martians would not need to attribute beliefs and desires to us if they wish to predict and explain what our bodies do, principle (N) entails that they would not be justified in attributing beliefs and desires to us. In contrast, we human beings do need these hypotheses, so (N) entails that we are justified in accepting them. It therefore becomes hard to see how mentalistic attributions can be objectively correct, since our warrant for accepting them seems to depend on our possessing certain sorts of cognitive limitations. Just as Laplace's thought experiment about his "demon" suggests that probability claims can't state objective truths if the universe is deterministic, so Nozick's thought experiment suggests that mentalistic ascriptions cannot state objective truths about human beings.

Many philosophers have agreed with Dennett (1987) that the right reply to this puzzle is to say that hypotheses can be justified on grounds other than their predictive utility. Granted, smart Martians would not need to attribute beliefs and desires if they wish to predict the behavior of bodies. However, mentalistic ascriptions are still needed because they provide better explanations than does a thorough-going physicalism.

In what sense are mentalistic explanations "better" than physicalistic explanations? Dennett (1987, pp. 25-27) says that mentalism allows one to identify patterns that would be invisible from a purely physicalistic point of view. He suggests that the Martians won't be able to see that a given behavior is the invariant outcome of a variety of initial conditions. Martians will explain why someone performs behavior B by saying that he was in physical state P_1, but won't realize that the same behavior would have resulted if he were in P_2 or P_3 or ... or P_n. However, I don't see why super-intelligent Martians would miss this fact. If they are smart enough to compute B from each of the initial conditions listed, why aren't they smart enough to see that B is invariant across this range of initial conditions?

A somewhat different way to argue that mentalistic explanations are better than physicalistic explanations is to claim that mentalism allows one to describe what physically different systems have in common, something that purely physical explanations are unable to do. Suppose two individuals each perform behavior B. A devotee of mentalism might suggest that this similarity is due to the fact that the two individuals were moved by the same belief/desire pair. In contrast, suppose that the Martians propose that the first individual performed the behavior because he had physical property P_1, while the second did so because she had P_2. If both hypotheses predict the behavior we observe, how are we to choose between them? If unification is a virtue of scientific hypotheses, then the mentalistic explanation has a virtue that the physicalistic explanation does not possess. It explains similar effects by postulating similar causes. This is said to justify our accepting the mentalistic explanation.

I won't dispute the claim that the mentalistic explanation is unified while the physicalistic explanation is not.[8] However, I am skeptical that this counts as <u>evidence</u> for the mentalistic hypothesis. If the unification provided by the mentalistic explanation is evidence in its favor, then the <u>dis</u>unification of the physicalistic explanation must count against it. But surely this isn't right in the context of the problem at hand. After all, we are asked to assume that the Martians are <u>smart</u> -- that they can identify the <u>true</u> physical state of the human bodies they inspect. The fact that the Martians see that the two individuals are different while the devotee of mentalism says that they are the same isn't evidence against the former <u>or</u> evidence in favor of the latter.[9]

The proper solution to the smart Martian problem, I suggest, is to reject the challenge of showing why the mentalistic hypothesis is more plausible than the physicalistic hypothesis. It is true that smart Martians wouldn't need the mentalistic ascription M and would be content to talk of P_1, P_2,..., P_n. But to extract from this a reason for doubting that people and other organisms have beliefs and desires -- and do so in as objective a fashion as you please -- is a mistake that comes from applying the proper procedure for testing rival causal hypotheses to pairs of hypotheses that are not rivals.

To see why, let's imagine that we want to explain why cigarette smokers get lung cancer more frequently than nonsmokers. Let's suppose that the frequencies are 1/100 and 1/1000, respectively. We consider two rival hypotheses. The first says that smokers are more apt to get lung cancer because smoking cigarettes involves inhaling smoke from burning paper. The second

hypothesis says that smoking is associated with lung cancer because smoking involves inhaling smoke from burning tobacco. The obvious way to test these hypotheses against each other is to find out how frequently people get lung cancer in each of four treatment groups, represented by the cells in the following 2-by-2 table. Suppose that the frequencies of lung cancer per 1000 people in a very large study look like this:

		inhaling smoke from burning paper	
		+	-
inhaling smoke from burning tobacco	+	10+e	10
	-	1+e	1

If e is small enough (given the sample size), one should accept the tobacco hypothesis and reject the paper hypothesis. A reasonable summary of this conclusion would be that the paper hypothesis isn't needed to explain why cigarette smokers get lung cancer more often than nonsmokers. The tobacco hypothesis suffices. On the other hand, if e turns out to be large enough, both hypotheses will be needed to explain the results. In this instance, the conclusion to draw is that cigarettes contribute to lung cancer by two causal pathways.

The smart Martian problem seems to have bite because it encourages us to see a mentalistic hypothesis and a physicalistic hypothesis as rivals, when, in fact, they are not. We are supposed to think that if Martians don't need to attribute beliefs and desires to explain behavior, then they are in the same position as the student of lung cancer who can explain the data just by invoking the tobacco hypothesis. Why bring in burning paper? Why invoke beliefs and desires? Both hypotheses seem to introduce fictions, though the fictions perhaps differ in their utility. This is a misleading analogy if ever there was one.[10]

With the data described above, it is true that one doesn't need the paper hypothesis to explain why smokers get cancer more often than nonsmokers. Notice, however, that this is a fact about the data, not about the cognitive powers of the investigator. The proper question is not whether we or Martians or anybody else needs to attribute beliefs and desires to explain behavior, but whether the data demand it.[11]

This confusion of an ad hominem question with one about the relation of observations to hypothesis is encouraged by an ambiguity in principle (N). We need to distinguish between prudential and evidential reasons -- a distinction that should be familiar from discussions of Pascal's wager (see, e.g., Mougin and Sober 1994). E may provide someone with a prudential

reason for believing H even though E is not an evidential reason supporting H. Notice how reference to an agent figures in the first of these relations but not in the second. In the present context, observed behavior may provide a human being with a prudential reason for attributing beliefs and desires to someone, while the same observations may fail to provide a Martian with a prudential reason for doing so. However, this says nothing as to whether the observations provide an evidential reason supporting the mentalistic attribution. The prudential needs of smart Martians and not-so-smart human beings have nothing to do with whether behavior provides evidence that people have beliefs and desires.

9. A law that says how probable a behavior is, given the presence of a multiply realizable supervening mental property, must be specific to one of the property's physical realizations, or the mental property must be causally complete with respect to the behavior.

Let M have two possible physical realizations (P_1 and P_2). The argument would be the same if there were more. Consider the following expansion of $Pr(B \mid M)$:

$$Pr(B \mid M) = Pr(B \mid M\&P_1)Pr(P_1 \mid M) + Pr(B \mid M\&P_2)Pr(P_2 \mid M) .$$

Laws must be time-translationally invariant. That is, if there is a law of the form "$Pr(B \mid M) = r$," then the value of r must be the same regardless of when t is. This is a plausible rendering of the idea that laws must be "universal" -- they must hold at all places and times (Earman 1986, Sober 1993b).

The values of $Pr(P_1 \mid M)$ and $Pr(P_2 \mid M)$ are not time-translationally invariant. In one temporal period, $Pr(P_1 \mid M)$ may be high; in another it may be low. This simply reflects the fact that the individuals at one time who have M might mostly have P_1, while the individuals at another time who have M might mostly have P_2. Given this, how can there be a law that relates B and M? This problem can be formulated in an older vocabulary. Laws are supposed to be necessary. Yet, $Pr(B \mid M)$ apparently depends on $Pr(P_1 \mid M)$ and $Pr(P_2 \mid M)$, which have their values only contingently. If so, how can a statement that assigns a value to $Pr(B \mid M)$ express a law?

There are two ways: (i) let $Pr(B \mid M) = Pr(B \mid M\&P_1) = Pr(B \mid M\&P_2)$; (ii) let there be two laws; one says "if P_1, then $Pr(B \mid M) = r_1$" while the other says "if P_2, then $Pr(B \mid M) = r_2$."[12] Option (i) says that the mental property M screens off its physical realizations from behavior B. It says that the mental is causally complete with respect to this behavior. Note that there is no contradiction between saying that physics is causally complete (with respect to all phenomena) and saying that the mental is causally complete with respect to a particular behavior. It could be the case that $Pr(B \mid M) = Pr(B \mid P_1) = Pr(B \mid P_2) = Pr(B \mid M\&P_1) = Pr(B \mid M\&P_2)$. Option (ii), on the other hand, says that a psychological law relating M to B must be specific to a given physical realization.

Although options (i) and (ii) are both possible in principle, option (i) isn't very plausible.

It is hard to believe that mental properties predict behavior so well that facts about the physical mechanisms in place are entirely irrelevant. The more extensive the P_i are in their specification of physical facts about an individual who exemplifies M, the less plausible it is to claim that M is causally complete with respect to B. So as to make the point vivid, let the different P_i include all the physical features an organism has when it has M. It now should be obvious that the mental property M will not be causally complete with respect to the behavior B. Even if your beliefs and desires this afternoon make it probable that you will watch a movie tonight, surely the probability of this behavior is additionally influenced by whether you will have a heart attack before the evening comes. Notice that there is nothing in the formula given above for the expansion of $Pr(B \mid M)$ that prohibits us from construing the P_i as providing a quite comprehensive profile of an individual's physical state. If option (i) is therefore implausible, we are forced to reject the idea that there are supervenient psychological laws that assign a value to $Pr(B \mid M)$ that generalize across systems that deploy different physical realizations for the mental properties in question.[13]

Papineau (1993, Chapter 2) has developed an evolutionary argument for expecting there to be psychological laws that generalize over different physical realizations. If P_1 evolves in one lineage as a proximate mechanism for causing organisms to perform behavior B in certain circumstances, and if P_2 evolves in another lineage for the same selective reason, then we should expect $Pr(B \mid P_1) = Pr(B \mid P_2)$. Block (forthcoming) endorses a similar thesis -- discovering the value of $Pr(B \mid P_1)$ in one lineage provides evidence for thinking that $(B \mid P_2)$ has that same value in another, if P_1 and P_2 each evolved because they cause the organism to perform behavior B. There are two reasons why Papineau's conclusion may fail. First, even if selection is the one and only process that determines which of the available phenotypes in the two ancestral populations evolves, there is no guarantee that the range of variation in the two populations is exactly the same. Second, there is the additional possibility that different nonselective factors may affect the two lineages. Convergent evolution is common, but I doubt that it can be counted on to insure the equality of probabilities we are here considering.[14] Block's formulation seems vulnerable to the same considerations. The observed value of $Pr(B \mid P_1)$ in one lineage is a good estimate of the unobserved value of $Pr(B \mid P_2)$ in another only to the extent that the two lineages have been shaped by the same evolutionary forces. Without any assurances on this matter, I don't see how the one value provides evidence about the other.[15]

Another way to try to save the idea of psychological laws that generalize over different physical realizations of a given mental property is to invoke the idea of ceteris paribus laws. If $Pr(B \mid M) = r_1$ when P_1 obtains and $Pr(B \mid M) = r_2$ when P_2 obtains, and r_1 and r_2 are both close to 1.0, why not say that the relevant psychological law is that "if M, then B, ceteris paribus?" This strikes me as an abuse of the ceteris paribus concept. If women smokers have a risk of lung cancer of 0.001 and men smokers have a risk of lung cancer of 0.002, these facts are not well summarized by saying that smokers don't get cancer, ceteris paribus. What are the other factors whose being "equal" (or absent) guarantee that smokers avoid cancer?

A better way to save the idea that laws in the special sciences can generalize over physically different realizations of a given supervenient property is to think of such laws as

providing value ranges, rather than point specifications, of probabilities. If $Pr(B \mid M\&P_1) = r_1$, $Pr(B \mid M\&P_2) = r_2$, and P_1 and P_2 are the only two physical realizations that M can have, then the proposition that $Pr(B \mid M)$ is between r_1 and r_2 will be time-translationally invariant. Laws in the special sciences may therefore have an approximate character in virtue of their generalizing over different physical realizations.

This way of retaining the idea of a higher-level law that generalizes over different physical realizations suggests that such laws are reducible, in at least one reasonable sense of that term. After all, we can explain the law that states the value range for $Pr(B \mid M)$ by saying that $Pr(B \mid P_1) = r_1$, that $Pr(B \mid P_2) = r_2$, and that P_1 and P_2 are the only physical realizations that M can have. A further consequence of this proposal is that a certain familiar way of arguing for the irreducible status of higher-level laws starts to look rather suspicious. Fodor (1975) and others have claimed that a mental property's list of possible physical realizations will be "open ended." However, if the list <u>is</u> open ended, how can we be confident that the point value or the value range we specify for $Pr(B \mid M)$ is correct? Although ignorance of the range of physical realizations may prevent us from reducing the higher-level statement to statements at lower level, it also raises the question of whether we are entitled to regard the higher-level statement as <u>true</u>.

Even though the present line of argument may make it look as if there is little hope of finding higher-level laws that generalize over different physical realizations, there is a way out of this problem that seems to be used quite extensively in evolutionary theory. Consider the concept of "fitness," the multiply realizable biological property <u>par excellence</u>. What do a fit zebra, a fit oak tree, and a fit bacterium have in common? Nothing much at the level of their physical properties, although all of them have the ability to survive and reproduce successfully in their environments. If fitness is multiply realizable, how can there be time-translationally invariant generalizations about fitness? This is not an empty question, because such generalizations exist in abundance. For example, R.A. Fisher (1930) established a proposition that he called the fundamental theorem of natural selection. This says that in populations of a certain type, the rate of increase in fitness at a time equals the additive genetic variance in fitness that exists at that time. This principle is true, no matter what time it is.

The trick that seems to have been used in evolutionary biology is to exploit the fact that the fitness of an organism <u>definitionally</u> entails facts about its probability of surviving and reproducing. If we focus just on the component of fitness that concerns an organism's chance of surviving, then to say that an organism has a (viability) fitness of 0.9 means that it has a probability of surviving from egg to adulthood of 0.9. Two physically different organisms might have the same value for this quantity:

` $Pr(o_1$ survives to adulthood $\mid o_1$ has a viability fitness of 0.9 and o_1 has $P_1) =$
$Pr(o_2$ survives to adulthood $\mid o_2$ has a viability fitness of 0.9 and o_2 has $P_2) = 0.9$.

Here fitness screens-off an organism's physical properties from its chance of surviving to adulthood (Brandon 1990; but see Sober and Wilson 1994 for some complications). This isn't

due to an empirical fact about fitness, but to its definition. There are time-translationally invariant laws[16] about fitness that generalize across different physical realizations of the supervenient property, and do so in conformity with option (i).[17] It remains to be seen whether cognitive science will pursue a similar strategy.

The argument of the present section concerns whether there are laws that assign a value, or a range of values, to probabilities of the form $Pr(B \mid M)$, where B is a behavior and M is an earlier mental state. Could psychological laws be formulated in a different way that escapes the dilemma of having to choose between options (i) and (ii)? Of course, psychological laws can relate one mental state to another; they need not advert to behavior. However, everything just said applies to laws that assign a value (or a range of values) to $Pr(M_2 \mid M_1)$, where M_1 is an earlier mental state and M_2 a later one.

A new issue is raised, however, if we consider claims to the effect that a particular mental property is a positive causal factor in the production of a behavior. If M raises the probability of B in each (qualitatively specified) background context, does this fact count as a law? The statement that M is a positive causal factor for B is a conjunction of inequalities between probabilities; each conjunct may hold true invariantly over time. If so, do we have a law? I see no reason to withhold the label. This suggestion for how "law" should be understood has the virtue of according better with what goes on in various areas of psychological research. Psychologists (and other social scientists) study effects; they try to determine whether a factor C increases the probability of E when other causal factors that impinge on E have been controlled for. Psychologists rarely try to determine the value or value range for probabilities of the form $Pr(B \mid M)$.

The idea of time-translational invariance, construed probabilistically, has an additional consequence that I want to mention. It concerns an intuition that Fodor (1975) has expressed about the concept of law, one that is crucial to his argument against reductionism. He says that even if it is a law that "if A then P" and a law that "if B then Q," it doesn't follow that "if A or B, then P or Q" is a law.[18] I find Fodor's reason for saying this obscure. He seems so sure that nature abhors a disjunction! Yet, the last of these conditionals is nomologically necessary, if the first two are. What is the magic ingredient that has somehow been lost? In a probabilistic setting, however, an analog of Fodor's claim is much less mysterious. Even if it is a law that "$Pr(P \mid A) = r_1$" and a law that "$Pr(Q \mid B) = r_2$," it doesn't follow that there is a law of the form "$Pr(P \text{ or } Q \mid A \text{ or } B) = r$." There is no need for a heavy-duty metaphysics of properties and natural kinds to explain why; the requirement that laws must be time-translationally invariant suffices.

10. Supervenience and the history of science

I now want to turn from metaphysics to epistemology. What justification could there be for the claim that mental properties supervene on physical properties? Surely this isn't an a priori truth. What is the evidence that the supervenience claim is correct?

One reasonable place to begin is with the successful physicalistic explanations that science has already developed. There are many biological properties and processes whose material bases used to be unknown. Yet, we now understand digestion, respiration, reproduction, and so on in terms of the physicalistic framework of molecular biology. It is only natural to expect the future to resemble the past. Although there is a great deal about psychological properties and processes that is currently not understood, it seems likely that neuroscience will do for memory, perception, sensation, and so on what molecular biology has already achieved for digestion, respiration, reproduction and so on. Of course, there is no deductive guarantee that physicalistic explanations of all psychological phenomena exist; but the best bet, given what has already been achieved in science, is that the general thesis of supervenience is correct.

Although I think there is something to this "inductive" argument from the past successes of physicalism, it repays closer scrutiny. First of all, it presupposes something like the assumption of random sampling. The argument treats psychological and biological properties as if they were balls in an urn, which scientists draw out at random and investigate. If all the balls drawn so far have been found to obey the supervenience thesis, then it is a reasonable induction to conclude that the rest of the balls in the urn are the same. The problem, however, is that scientists don't choose their problems by sampling at random. Scientists choose problems by deciding which problems are amenable to the techniques they have at hand. If so, the fact that physicalism worked fine as a framework for understanding respiration and digestion doesn't settle whether physicalism is likely to succeed as a framework for understanding consciousness.[19]

This may suggest that one should be completely agnostic on the issue of whether supervenience is right; however, I don't think that supervenience and its negation are on an epistemic par. But before I say why, I want to explore an additional gap in the argument from the history of science. Curiously, this further limitation in the argument helps show why the supervenience claim has a special status not enjoyed by its negation.

Let's consider in more detail a phenomenon that now is regarded as uncontroversially physicalistic in character. Suppose that respiration is such an example. What sort of evidence could there be for the claim that respiratory properties supervene on physical properties? One might be tempted to answer by citing the large body of well-attested theory that describes the physical bases of various respiratory subprocesses. However, we need to consider more closely what type of evidence these physicalistic theories enjoy. It is easy enough to see how observations could show that this or that physical property contributes to, or detracts from, the ability to breathe. However, what sort of evidence could there be for the claim that physical properties exhaust what is relevant -- that the physical properties present at time t determine an organism's ability to breathe at time t? Don't the data always leave at least some room for the possibility that there exists a nonphysical property that makes a tiny difference in whether someone can breathe even after all physical properties are held fixed?

The more general question is this: Exactly what sort of evidence can there be for the supervenience claim (S) that $Pr(M \mid P) = 1$, where M is some biological or psychological property

that a system has at a given time, and P is the totality of its physical properties at that time? How do we know that the value of the conditional probability isn't something less than unity? Notice that even if there were 100,000 cases in which M occurred when P did, and no observed case in which M failed to occur when P was present, this still would be consistent with the hypothesis

(H) $Pr(M \mid P) = 0.99999999999999999999999999999999999999999999999$.

Evidently, the data don't permit one to reject (H). Why, then, is it legitimate to interpret even the most successful instances of physicalistic science as vindicating a specific supervenience claim?

The physicalist might be inclined to reply that it is part of the scientific method to look for physical causes. If we actually had evidence that a list of the physical properties exemplified by an object at a given time is insufficient to determine whether the nonphysical property M is present at that time, then we should endeavor to discover what further physical properties have to be added to our list. To summarily conclude that M fails to supervene on <u>all</u> physical properties is tantamount to closing down inquiry.

These methodological remarks are well taken, but they don't provide the justification we seek. The question is not whether one should look for physical causes that provide a sufficient supervenience base, but why one should believe that such a supervenience base in fact exists. Bear in mind that we are at present not considering whether the supervenience thesis is true for some poorly understood higher-level property (e.g., consciousness); rather, we are considering a biological or psychological property that is supposed to be a paradigm case of triumphant physicalistic explanation. How could observation justify the acceptance of <u>any</u> supervenience claim?

This focus on the issue of how observations should be interpreted is worthwhile in its own right, but it is especially important in view of the fact that we at present have no compelling "top/down" argument in favor of supervenience. For example, we saw earlier in this paper that (S) is not entailed by (CCP), or by the conjunction of (CCP) and (PDM). Supervenience is a <u>strong</u> physicalistic claim. Perhaps some new theoretical argument can be supplied, but as things now stand I know of none.

<u>11. Supervenience as a null hypothesis</u>

How would a scientist go about testing a hypothesis of the form $Pr(M \mid P) = 1.0$? It isn't essential to think of M and P as mental and physical properties; they could be any two properties you please. One advantage of understanding the supervenience thesis as advancing a probabilistic claim is that it can be assimilated to quite routine problems of hypothesis testing in science.

The first point that needs to be made is that a scientist asked to test (S) would not dream

of testing the hypothesis (S) against the hypothesis (H). More likely, the problem would be formulated by contrasting the supervenience claim with its negation, namely with the claim

(not-S) $Pr(M \mid P) = \alpha$, where $\alpha < 1.0$.

Whereas the supervenience thesis would be termed a "simple" statistical hypothesis, its negation is "composite." Notice that (not-S) does not predict how often P objects should have M.

It might be asked how the simple hypothesis (S) can be tested against the composite hypothesis (not-S), if the latter doesn't tell you what to expect. The usual procedure is to ignore the composite hypothesis and to ask whether the observations suffice for one to reject the simple hypothesis. If one looked at 100 individuals who have P, and found that all had M, one would not reject (S); if the fraction turned out to be less than 100%, one would reject the supervenience claim.[20] Understood in this way, the supervenience thesis has the status of a "null" hypothesis; it is regarded as innocent until proven guilty. Observing 100% Ms among a sample of Ps also is consistent with (not-S), but that is somehow not taken to be relevant.

I've just described what I take to be the practice of scientists who wish to test a hypothesis of the form given by (S). The theory that underwrites this practice is another matter. I suggest that the standard Neyman-Pearson approach doesn't fully account for what is going on here. The problem that I want to highlight concerns why (S) is regarded as innocent until proven guilty. My question is not why one should favor (S) over (H). Let's suppose that the problem of interest is the comparison of (S) with (not-S). Neyman-Pearson theory treats (S) as the null hypothesis because it doesn't have the resources to test (not-S) at all. Since (not-S) makes no predictions about frequencies, the approach can't say what it would take to reject that hypothesis. Fair enough, but why is that a reason to regard (S) as innocent until proven guilty? Neyman-Pearson theory offers no justification for imposing this asymmetry.

Bayesianism, which is a very different approach to the problem, doesn't do any better. If one observes a thousand P objects and all turn out to be M, then the posterior probabilities of (S) and (not-S), relative to these data, obey the following formula:

$$Pr(S \mid Data) > Pr(\text{not-}S \mid Data) \text{ if and only if}$$
$$Pr(Data \mid S)Pr(S) > Pr(Data \mid \text{not-}S)Pr(\text{not-}S).$$

$Pr(Data \mid S) = 1$, of course, but what value should one assign to $Pr(Data \mid \text{not-}S)$? The composite hypothesis (not-S) is a disjunction; different disjuncts (D_i) confer different probabilities on the data. The expression $Pr(Data \mid \text{not-}S)$ expands into a summation:

$$Pr(Data \mid \text{not-}S) = \sum Pr(Data \mid D_i)Pr(D_i \mid \text{not-}S).$$

What values should be assigned to probabilities of the form $Pr(D_i \mid \text{not-}S)$? If $Pr(M \mid P) < 1.0$, how probable is it that $Pr(M \mid P) = 0.1$, that it equals 0.2, and so on? It is hard to see how the

assignment of values to these second-order probabilities can be justified objectively.

The same sort of problem arises for the prior probabilities Pr(S) and Pr(not-S). It is not clear why (S) should be assigned a higher prior probability than (not-S). It also isn't clear that they should be assigned equal priors. In fact, any proposed assignment seems to be more a confession of subjective confidence than a report on anything objective. It isn't that Bayesianism says that supervenience is on an evidential par with its negation, but that no determinate answer seems to be forthcoming concerning what one ought to think.[21]

12. The Akaike framework

There is a third approach to this problem, one that I think makes more sense than either Neyman-Pearson testing or Bayesianism. H. Akaike (1973) is a Japanese statistician who developed a criterion that judges the predictive accuracy that models can be expected to have (Forster and Sober 1994). Both these italicized expressions require explanation.

A model is an equation that has some number of adjustable parameters. For example, consider the hypothesis that says that y is a linear function of x:

(LIN) $y = a + bx$.

(LIN) is the set of all straight lines. Fix values for the parameters a and b, and a unique straight line is determined. Another model says that the relation of x and y is parabolic:

(PAR) $y = a + bx + cx^2$.

Again, fix the values of a, b, and c, and a unique parabola is identified.

The problem that Akaike addressed is that of model selection; the goal is to find the model that has the best estimated degree of predictive accuracy. One begins with a set of data -- points in the <x,y> plane, say -- and uses these data to select a model. The question is how one is to use the data at hand to choose a model that will do the best job of predicting new data. For example, if one uses the model (LIN) in this problem, one will find the straight line that comes closest to the existing data,[22] and then use this straight line, which I'll call L(LIN), to predict new data. If the true underlying curve is in fact a straight line, then LIN will probably do well in this task; on the other hand, if the true curve is in fact extremely nonlinear, L(LIN) will probably do a poor job predicting new data.

Before I describe Akaike's solution to this problem, I want to mention a phenomenon that empirical scientists encounter when they fit a model to data and then use the fitted model to predict new data. It is easy to get a model that fits the old data as well as you please. By making a model sufficiently complex -- by increasing the number of adjustable parameters it contains --

you can fit the data to any desired degree of accuracy. However, scientists find that when they make their models complicated to achieve good fit to old data, the resulting fitted model often does a poor job of predicting new data. When this happens, scientists say that the model overfit the old data. The old data, so to speak, contain both signal and noise; a very complicated model mistakes noise for signal. This is the workaday understanding that many scientists have of why simplicity is relevant to the task of model selection. Simpler models often fit the old data worse than more complicated models, but they often do a better job of predicting new data. Since the goal is to predict new data, one wants to give some weight to simplicity.

This bit of phenomenology is important philosophically because it helps demystify the role of simplicity in scientific inference. However, it leaves several questions unanswered. If simplicity matters in model selection, how much does it matter? There must be some principled trade-off between fit to old data and simplicity. What is the right rate of exchange? Most fundamentally, we need to ask why scientists so often have the experience just described. Is it due to some basic metaphysical fact concerning the simplicity of nature?

The goal is to estimate how predictively accurate a model will be. Certain facts about the model we are considering are known. We know how well the best fitting member of the model fits the data at hand; we also know how simple the model is (we just count the number of adjustable parameters it contains). How can we use these known quantities to estimate how well the model will do in predicting new data that we haven't even seen? After all, the success that the model will enjoy in this task depends on what the true underlying curve is, and that curve is, by hypothesis, unknown.

Akaike was able to prove that the following quantity is an unbiased estimate of the predictive accuracy of a model M:

$$\text{log-Pr}[\text{Data} \mid L(M)] - k\sigma^2 + \text{constant}.$$

Here "log-Pr[Data | L(M)]" denotes the logarithm of the probability that the best-fitting member of M confers on the data, k is the number of adjustable parameters, and σ^2 characterizes how much error there is likely to be in the observations. Akaike's result means that there are two factors that contribute to a model's expected degree of predictive accuracy -- its fit to old data and its simplicity. When we compare two models -- (LIN) and (PAR), for example -- the constant term disappears. So, if (LIN) fits the data about as well as (PAR) does, Akaike's theorem tells us that we should prefer (LIN); its greater simplicity means that it can be expected to have a higher degree of predictive accuracy.

There are several further issues that are important to this problem, including the fact that other criteria of model selection have been proposed that give simplicity a different weight from the one provided by Akaike's theorem. However, these details won't matter to the lesson I want to draw in connection with the problem of supervenience. The simple fact of interest about the comparison of (S) and (not-S) is that the supervenience claim contains no adjustable parameters,

whereas its negation contains one (namely α). This means that (S) is simpler in a sense that matters to the task of model selection.[23]

Suppose we examine a large number of individuals and check whether each has P and whether each has M. We detect the presence of P by some complex neurophysiological test and ascertain whether subjects are in mental state M by recording their verbal reports. For Akaike's framework to apply to this problem, we have to assume that these observations are subject to error -- not an implausible assumption, by any means. If so, the supervenience thesis does not predict that 100% of the individuals we label as having P are also labeled as having M. Depending on the size of the error probabilities, the frequency predicted by (S) will be something less; suppose it is 98%. We then see what the frequency in the data actually is. Suppose it is 96%. We then find which member of the family (not-S) maximizes the probability of the observations; L(not-S) will set the parameter $\alpha = 0.96$. Notice that the supervenience hypothesis fits the data less well than L(not-S); the question is whether the greater simplicity of (S) compensates for this deficiency. Notice that (not-S) can't do a worse job of fitting the data than (S). However, the question isn't which hypothesis fits the old data better, but which will do better in predicting new data. In this context, the fact that (S) is simpler than (not-S) can lead us to prefer (S).

13. The significance of the simplicity of the supervenience thesis

The Akaike framework helps explain why the hypothesis of supervenience and its negation are not on an epistemic par. However, it is important not to exaggerate what we are entitled to conclude. The Akaike framework applies when one considers a specific property M and a specific property (or conjunction of properties) P. It may be that M fails to supervene on P, but does supervene on some other constellation of physical properties as yet undiscovered. Perhaps one should retain the general thesis of supervenience even after numerous attempts to locate the supervenience basis of M have failed. This may make sense, but I am not suggesting that the Akaike framework explains why. Rather, what the Akaike framework addresses is how observational data about the frequency of M individuals among P individuals ought to be interpreted. It helps clarify why successful cases of physicalistic science are rightly interpreted as vindicating physicalism. Physicalistic models have fit the data well enough that it would be a mistake to adopt more complex models that deny that the supervenience relation obtains. It makes sense to claim that breathing and digestion supervene on physical processes, even though the data also are consistent with the hypothesis that supervenience is false. There is more to model selection than fitting the data.[24]

Hard-line physicalists may find this vindication of supervenience too limited. Perhaps they feel sure that consciousness must fall into line within the physicalistic world picture, just as respiration and digestion have done. This may be a perfectly good working hypothesis, in that it structures and directs research in a productive direction. In fact, it may be the only serious working hypothesis, if the denial of supervenience is too flabby an idea to be heuristically useful.

However, these methodological properties of the general supervenience thesis need to be separated from the question of whether and why we are entitled to believe that it is more plausible than its negation. The Akaike framework shows why simplicity matters in scientific inference, but it doesn't show that simplicity considerations are a substitute for looking at data. Rather, simplicity is one consideration in the task of model selection and fit-to-data is another. It would be an abuse of the Akaike framework to think that it shows that consciousness must supervene on physical properties, independent of any data about how states of consciousness and specific physical states are in fact associated. However, once data of this sort are obtained, Akaike helps us understand why the supervenience thesis is preferable to its negation, ceteris paribus. Whether other factors really are equal is an empirical question that can be judged only by looking at data.

It is interesting that the issues relevant to choosing between (S) and (not-S) are the same, regardless of how M and P are interpreted. We have thought of M as a nonphysical property that a system might exhibit at a given time and P as the list of all the physical properties that the system exhibits at that time. However, the same analysis would apply if M and P were two physical properties exemplified at the same time, or at different times. The hypothesis of diachronic determinism has a special inferential status that the hypothesis of indeterminism does not possess; this asymmetry precisely parallels the one that obtains between the hypothesis of synchronic supervenience and its negation.

This parallel between the synchronic thesis of supervenience and the diachronic thesis of causal determinism is instructive. Kant claimed that determinism was necessary for the possibility of experience. This, we now see, was an exaggeration, whose kernel of truth is that determinism is a null hypothesis. Perhaps future philosophers will come to view the present high regard that supervenience enjoys as similarly exaggerated. Not that I expect supervenience to turn out to be false. However, the insistence that it must be true goes considerably beyond what current evidence and argument warrant.

14. Concluding comment

The supervenience thesis in its general form has the quantifier order $(\forall)(\exists)$. It says:

> For each mental or biological property M, if a system possesses M at a given time, then there exists a set P of physical properties that the system also possesses at that time such that $Pr(M \mid P) = 1.0$.[25]

I have explored the metaphysics and epistemology of instances of this thesis; these are the result of selecting a particular property M and a particular set P of physical properties. The metaphysical side of the inquiry involved examining the relationship of instances of the general supervenience claim to other metaphysical claims. The epistemology involved asking how evidence could be brought to bear on the claim that $Pr(M \mid P) = 1.0$. It is striking how a probabilistic point of view helps clarify this range of issues. It is possible, of course, that there is

more to the thesis of physicalism than is dreamed of in the simple assignment of value to a conditional probability. However, if this probabilistic formulation captures part of what physicalism involves, then the probabilistic point of view is a point of view well worth taking.

References

Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle." In B. Petrov and F. Csaki (eds.), Second International Symposium on Information Theory. Budapest: Akaemiai Kiado, 267-281.

Block, N. (forthcoming): "Anti-reductionism slaps back." Philosophical Perspectives.

Brandon, R. (1990): Adaptation and Environment. Princeton: Princeton University Press.

Brandt, R. and Kim, J. (1967): "The logic of the identity theory." Journal of Philosophy 64: 515-537.

Causey, R. (1977): The Unity of Science. Dordrecht: Reidel.

Crane, T. and Mellor, H. (1990): "There is no question of physicalism." Mind 99: 185-206.

Davidson, D. (1970): "Mental events." In L. Foster and J. Swanson (eds.), Experience and Theory. London: Duckworth. Reprinted in Essays on Action and Events. Oxford: Oxford University Press, 1980, 207-228.

Dennett, D. (1987): "True believers." In The Intentional Stance. Cambridge: MIT Press.

Earman, J. (1986): A Primer on Determinism. Dordrecht: Reidel.

Eells, E. (1991): Probabilistic Causality. Cambridge: Cambridge University Press.

Enç, B. (1986): "Essentialism without individual essences -- causation, kinds, supervenience, and restricted identities." In P. French, T. Uehling, and H. Wettstein (eds.), Midwest Studies in Philosophy, vol. 11. Minneapolis: University of Minnesota Press. 403-426.

Fisher, R. (1930): The Genetical Theory of Natural Selection. Oxford: Oxford University Press.

Fodor, J. (1975): The Language of Thought. New York: Crowell.

Fodor, J. (1987): Psychosemantics. Cambridge: MIT press.

Forster, M. and Sober, E. (1994): "How to tell when simpler, more unified, or less ad hoc

theories will provide more accurate predictions." British Journal for the Philosophy of Science 45: 1-35.

Hellman, G. (1985): "Determination and logical truth." Journal of Philosophy 82: 607-616.

Hempel, C. (1969): "Reduction -- Ontological and Linguistic Facts." In S. Morgenbesser, P. Suppes, and M. White (eds.), Philosophy, Science, and Method. New York: St. Martin's.

Horgan, T. (1993): "From supervenience to superdupervenience -- meeting the demands of a material world." Mind 102: 555-586.

Kim, J. (1989a): "The myth of non-reductive materialism." Proceedings of the American Philosophical Association 63: 31-47.

Kim, J. (1989b): "Mechanism, purpose, and explanatory exclusion." In Philosophical Perspectives. Vol. 3. J. Tomberlin (ed.), Atascadero, California: Ridgeview Press. 77-108.

Kim, J. (1990): "Explanatory exclusion and the problem of mental causation." In E. Villanueva (ed.), Information, Semantics, and Epistemology. Oxford: Blackwell, 36-56.

Loewer, B. (1995): "An argument for strong supervenience." In. E. Savellos and U. Yalcin (eds.), Supervenience -- New Essays. Cambridge: Cambridge University Press, 218-225.

McLaughlin, B. (1992): "The rise and fall of British emergentism." In A. Beckermann, H. Flohr, and J. Kim (eds.), Emergence or Reduction? Berlin: de Gruyter, 49-93.

Mougin, G. and Sober, E. (1994): "Betting Against Pascal's Wager." Nous 28: 382-395.

Orzack, S. and Parker, G. (1990): "Genetic variation for sex ratio traits within a natural population of a parasitic wasp, Nasonia vitripennis." Genetics 124: 373-384.

Papineau, D. (1990) "Why supervenience?" Analysis 50: 66-71.

Papineau, D. (1991): "The reason why -- response to Crane." Analysis 51: 37-40.

Papineau, D. (1993): Philosophical Naturalism. Oxford: Blackwells.

Papineau, D. (1995): "Arguments for supervenience and physical realization." In E. Savellos and U. Yalcin (eds.), Supervenience -- New Essays. Cambridge: Cambridge University Press. 226-243.

Reichenbach, H. (1956): The Direction of Time. Berkeley: University of California Press.

Rosenberg, A. (1994): Instrumental Biology or the Disunity of Science. Chicago: University of Chicago Press.

Segal, G. and Sober, E. (1991): "The causal efficacy of content." Philosophical Studies 62: 155-184.

Smart, J. (1959): "Sensations and brain processes." Philosophical Review 68: 141-156.

Sober, E. (1984): The Nature of Selection. Cambridge: MIT Press.

Sober, E. (1993a): Philosophy of Biology. Boulder: Westview Press.

Sober, E. (1993b): "Temporally oriented laws." Synthese 94: 171-189. Reprinted in From a Biological Point of View. Cambridge: Cambridge University Press, 1994, pp. 233-252.

Sober, E. (1996a): "Evolution and optimality -- feathers, bowling balls, and the thesis of adaptationism." Council on Philosophic Exchange Annual 26: 40-57.

Sober, E. (1996b): "Parsimony and predictive equivalence." Erkenntnis 44: 167-197.

Sober, E. (forthcoming): "Two outbreaks of lawlessness in recent philosophy of biology." Philosophy of Science //: ///-///.

Sober, E. and Wilson, D. (1994): "A critical review of philosophical discussion of the units of selection problem." Philosophy of Science 61: 534-555.

Notes

*. I thank Martin Barrett, Tom Bontly, Tim Crane, Ellery Eells, Berent Enç, Branden Fitelson, Malcolm Forster, Peter Godfrey-Smith, Leslie Graves, Daniel Hausman, Terry Horgan, Jaegwon Kim, Barry Loewer, Brian McLaughlin, Hugh Mellor, Greg Mougin, David Papineau, John Post, Alan Sidelle, Larry Shapiro, and Dennis Stampe for comments on an earlier draft.

1. My division of metaphysical from epistemological issues is somewhat artificial, in that several of the authors I'll consider have tried to settle the epistemological status of the supervenience thesis by showing that it bears implication relations to other metaphysical claims that we have independent reasons to accept or reject.

2. I don't deny that there may be further asymmetries between mental and physical properties. For example, Enç (1986) has discussed the intuition that a system's physical properties at a time "determine" its mental properties at that time, but not conversely, even in a domain in which there is a one-to-one mapping. I have nothing to say about how this intuition might be clarified, or on whether it is ultimately correct.

3. This probabilistic formulation of the supervenience thesis should be understood as a quantified statement: For any system, and for any mental property M, if the system has M at a given time, then there exists a set of physical properties P such that the system has P at that time and Pr(M | P) = 1.0. Notice that no mention of "completeness" is needed here, since if Pr(M | P) = 1.0, then Pr(M | P&Q) = 1.0 also. The idea of listing "all" the physical features of the system isn't essential. Note, in addition, that the conditional used in the supervenience thesis, like those in the other doctrines I'll discuss, should not be understood truth-functionally. It isn't enough for the supervenience thesis to be true that all actual objects that differ mentally also happen to differ physically.

4. As was true for the supervenience thesis, (CCP) is a quantified statement. It has the quantifier order $(\forall)(\exists)(\forall)$: For every property B, if the system has B at time t+dt, then there exists a set of physical properties P that the system has at time t, such that for all other properties M that the system has at time t, Pr(B | P) = Pr(B | P&M).

5. With this Reichenbachian analogy in hand, I can explain in more detail why I characterized (PDM) as I did. It would be a mistake to define the detectability of M by B as Pr(B | M&P)≠Pr(B | not-M&P). This is the definition of emergent causation (EC) and is <u>incompatible</u> with (CCP). These metaphysical issues aside, the definition given of detectability is a natural one; today's barometer reading is a detector (an indicator) of tomorrow's weather simply because they are correlated; the fact that today's weather renders them conditionally probabilistically independent does not undermine this fact.

6. The emergentists whom McLaughlin (1992) discusses sometimes suggest that additivity is required for properties of the whole to be "predictable" from properties of the parts "taken in isolation." However, this is a mistake. Without knowing the composition principle, you can't make a prediction, regardless of whether the relationship is additive in fact. And if you <u>do</u> know the compositional principle, you can make the prediction, whether or not the relationship is additive. Of course, knowing <u>merely</u> that additivity fails (i.e., that $c \neq 0$) means that the upper-left cell cannot be predicted from the entries in the other three, whereas knowing that additivity obtains (i.e., that $c = 0$) permits this prediction to be made.

7. The policy that Eells (1991) recommends allows one to show that (CCP) and (CEM) together entail (S), and, equivalently, that (CCP) and (not-S) together entail (not-CEM). A different policy on what must be included in background contexts could have different consequences.

8. Both individuals fall under the disjunctive predicate "P_1-or-P_2." To avoid having to concede that there is a unified physicalistic explanation of why the two individuals both perform the behavior, one needs to maintain that this disjunctive predicate doesn't pick out a physical property.

9. There are circumstances in which a unified hypothesis is more plausible than a disunified competing hypothesis, but this is not one of them. The confirmational framework described later in this paper helps explain when and why unification is epistemologically significant; see Forster

and Sober (1995) for details.

10. If we modify Nozick's problem a bit, we can formulate a problem in which mentalistic and physicalistic hypotheses do compete. Let's imagine that Martians are not so smart. Rather than assuming that they are using the true laws of physics and that they ascertain the states of the elementary particles in a person's body with perfect accuracy, let's suppose that their physicalistic description of these particles is subject to error and that their model for how such data should be used to predict behavior may or may not be true. If their model contains a large number of parameters whose values are inferred from observations, and the belief/desire model contains a much smaller number of such parameters, then it can turn out that the complexity of the physicalistic model provides a reason to expect it to be less predictively accurate. The Akaike framework described in the second half of this paper is relevant to understanding how this can happen.

11. Fodor (1987, p. 9) observes that "even if psychology were dispensable in principle, that would be no argument for dispensing with it. (Perhaps geology is dispensable in principle; every river is a physical object after all...)... What's relevant to whether common sense psychology is worth defending is its dispensability in fact." Fodor is right that the smart Martian problem applies to rivers just as much as it does to beliefs and desires. However, the resolution of the problem, whatever its target, does not consist in the fact that we are cognitively limited and therefore need to talk about rivers and mental states.

12. Actually, the situation is a bit more complicated. The two options listed are exhaustive if the P_i are not just minimally sufficient supervenience bases for M but are sufficiently inclusive that probabilities of the form $Pr(B \mid M\&P_i)$ are time-translationally invariant. However, if probabilities of the form $Pr(B \mid M\&P_i)$ are not time-translationally invariant, then the third way for $Pr(B \mid M)$ to be time translationally invariant is for the three logically independent quantities $Pr(B \mid M\&P_1)$, $Pr(B \mid M\&P_2)$, and $Pr(P_1 \mid M)$ to have their evolution coordinated, so that $Pr(B \mid M)$ always has the same value through time even though these component probabilities do not. I view this third option as too implausible to be worth exploring.

13. Of course, the methodological notion of autonomy -- that progress can be made in different parts of cognitive science without thinking about neural realization -- is left open.

14. Although evolutionary theory provides little reason for predicting that $Pr(B \mid P_1) = Pr(B \mid P_2)$, if this equality is discovered to obtain, the theory may be able to explain why by appeal to the selection hypothesis that Papineau describes.

15. Empirical and theoretical work in sex-ratio evolution provides an instructive analogy. Evolutionary theory describes the optimal mix of sons and daughters a parent should produce as a function of the breeding structure of the population. The optimal value in one species may therefore differ from the optimal value in another. That is, if B is the behavior of giving birth to a male, and P_1 and P_2 are the physical mechanisms for sex ratio determination that evolve in two different species, then it may happen that the optimal value for $Pr(B \mid P_1)$ differs from the optimal

value for $Pr(B \mid P_2)$. It is an empirical question how close to the optimal the individuals in a given species come on average; it also is an empirical question whether the individuals in a species depart from the optimal sex ratio to the same extent. Orzack and Parker (1990) found significant individual variation in the species of parasitic wasp they studied.

16. I call Fisher's fundamental theorem, when properly stated, a law, even though it is a mathematical truth. I discuss the idea that laws need not be empirical in Sober (forthcoming).

17. In Sober (1984, 1993a, forthcoming), I asked why generalizations in evolutionary biology seem, upon clarification, to either be accidental, or to be mathematical truths. Why aren't there empirical laws in this subject? The present point about time-translational invariance may provide part of the answer. Rosenberg (1994) also suggests, but by way of a different argument, that the supervenience of biological on physical properties explains biology's lack of empirical laws.

18. Here Fodor is following the lead of Davidson (1970), who says that although it is a law that "all emeralds are green" and the same is true of "all rubies are red," it isn't a law that "all emerubies are gred."

19. Even though scientists didn't select their problems at random, there was no a priori guarantee that physicalistic accounts of respiration, digestion, and so on would be obtained. These past successes can't be dismissed as irrelevant to the question of what the future will bring.

20. If observation is assumed to be subject to error, one should reject (S) precisely when the fraction of observed Ps that are M is significantly less than 100%.

21. This problem does not disappear by interpreting probabilities as subjective degrees of belief. Although there is no logical prohibition against assigning (S) a prior probability that is equal to or higher than the prior assigned to (not-S), it should give Bayesians pause to do so. If I drop a dart on a line a mile long, isn't it less probable that the dart will fall at the line's precise beginning than that it will fall somewhere else on the line?

22. The goodness of fit of a curve to a data set is standardly measured by computing its sum-of-squares. For each data point $\langle x,y \rangle$, one computes its distance from $\langle x,y_c \rangle$, which is the point on the curve that has the same x value. One squares this distance and then computes the sum of squared distances for all data points. Sum of squares is a valid measure of the curve's likelihood, meaning the probability that the curve confers on the data, if there is a symmetric error distribution. It is likelihood that is the fundamental measure of goodness-of-fit here; this is why I've called the straight line that fits the data best L(LIN).

23. It also is true that (H) is simpler than (not-H) in a sense that matters to the Akaike framework. There is no getting around the fact that the answer you get depends on the question you ask. It is perhaps unreasonable to expect the Akaike framework, or any other, to tell you whether you should test (S) against its negation rather than (H) against its negation.

24.Smart (1959) and Brandt and Kim (1967) defended the mind/brain identity theory by appeal to parsimony; see Sober (1996b) for discussion of their argument in the context of the Akaike framework.

25. It is interesting that another much discussed ism -- adaptationism -- has a similar quantifier structure. Rather similar methodological and evidential considerations bear on it. See Sober (1993a, 1996a) for discussion.