Journal of Psychology and Theology
2001, Vol. 30, No. 1, 41-50

Evolutionary Explanation and Consciousness

Steven Horst
Wesleyan University

If there is a received orthodoxy in contemporary philosophy of mind and cognitive science, it is that all features of the mind, including meaning, action, and consciousness, can and perhaps must be *naturalized* (i.e., accommodated within the framework of the world of nature as understood by the natural sciences). At the same time, just about everyone working in philosophy of mind has realized for at least a decade that there are substantial problems with this enterprise of naturalizing the mind, in that there are features of the mind that do not seem to lend themselves to naturalistic explanation: in particular, meaning, consciousness, and free will. For naturalists (who comprise the majority of the field at the moment), this is seen as an urgent problem: they feel that we *must* naturalize the mind, and yet it looks as though we do not know how to do so. Oddly, even many of the writers who have forcefully argued that the mental cannot be reduced to the physical, or cannot be explained in evolutionary terms, still call themselves "naturalists," even when they have no concrete naturalization of the mind to offer. A few other anti-naturalists and I, on the other hand, have tried to argue that the problems the naturalist faces are abiding and principled problems, and not merely a symptom of a current lack of development of psychology or neuroscience. In this article for the special issue, I shall attempt to do three things. First, I shall attempt to describe the current situation in philosophy of mind and cognitive science with respect to one special facet of the mind: consciousness. I shall try to explain, in a very abbreviated form, the historical factors that have led to the popularity of reductionist forms of naturalism, and then summarize several influential arguments to the effect that consciousness cannot be reduced to neuroscience or physics. I shall then differentiate two very different strands of this conversation: one which is about *explanation* (i.e., physical science cannot explain consciousness), and another which is about *metaphysics* (i.e., physical facts are not enough to determine facts about consciousness). Finally, I shall explain why I think that naturalizers of the mind can look for no solace from evolutionary explanation if—as appears to be the case—attempts at reductive explanation of the mind in terms of physics or neuroscience should continue to fail: in brief, because evolutionary explanation depends upon the fulfillment of a promissory note which only reductive explanation could make good on. It is thus only in the final section of this article that I shall approach the uniting theme of the special issue. However, to understand the status of evolutionary explanation in philosophy of mind, it is in my opinion necessary to see where it fits within a broader framework of naturalizing the mind.

Correspondence concerning this article may be addressed to Steven Horst, Department of Philosophy, Wesleyan University, Middletown, CT 06459.

Naturalizing Consciousness and the Explanatory Gap

Over the past twenty years, a large proportion of the work done in philosophy of mind has been framed in terms of the enterprise of *naturalizing* the mind, or accommodating it within the framework of the world of nature as understood by the natural sciences. On the surface, at least, naturalism appears to be very close to a consensus view in philosophy of mind. If you read books written in philosophy of mind in the last twenty years, you will find a growing trend towards describing one's own project as an attempt to "naturalize" the mind, and indeed to cast one's discussion within the assumption that what everyone is looking for is a "naturalistic" theory of the mind. One might even view naturalism as the prevailing trend of the entire twentieth century. For example, Jerry Fodor writes, "Here, then are the ground rules. I want a *naturalized* theory of meaning; a theory that articulates, in nonsemantic and nonintentional terms, sufficient conditions for one bit of the world to *be about* (to express, represent, or be true of) another bit" (1987, p. 98).

Why should one want such a theory? Because of the fear that mental states are somehow called into doubt if they cannot be naturalized. "The deepest motivation for intentional irrealism [i.e., the view that mental states are not real] derives . . . from a certain ontological intuition: that there is no place for intentional categories in a physicalistic view of the world; that the intentional can't be *naturalized*" (Fodor, 1987, p. 97). Fodor (1987) also states:

> It's hard to see . . . how one can be a Realist about intentionality without also being, to some extent or other, a Reductionist. If the semantic and the intentional are real properties of things, it must be in virtue of their identity with (or maybe their supervenience on?) properties that are themselves neither intentional nor semantic. If aboutness is real, it must be really something else. (p. 98)

Steve Stich and Steve Laurence (1994) cite a second and related concern: "Another, rather different concern is that if naturalization fails, then there could be no serious *science* of intentional psychology because there could be no *laws* that invoke intentional terms or intentional properties" (p. 161; note that Stich & Laurence are arguing against this catastrophic view of the failure of naturalism). Even David Chalmers, who rejects physicalism, sets a kind of naturalism as a constraint for his theory: "The third constraint is that I take consciousness to be a natural phenomenon, falling under the sway of natural laws" (Chalmers, 1996, p. xiii). The tone of these quotes is not so much "*Can* the mind be naturalized" but "*How* shall we naturalize the mind?" Naturalism is thus widely treated as a consensus view in philosophy of mind, and those who are not themselves of a naturalistic bent may well feel that they have not been invited to the conversation.

However, in fact the story is much more complex. For while a majority of people in philosophy of mind today would *label* themselves as naturalists, there seems to be little consensus about what this label in fact *means*. This fact has been widely noted, and has been remarked upon for perhaps half a century now. The philosopher of science Ernest Nagel, in his 1955 presidential address to the American Philosophical Association, noted that "the number of distinguishable doctrines for which the word 'naturalism' has been a counter in the history of thought is notorious" (Nagel, 1956, p. 3). In their introduction to the anthology, *Naturalism: A Critical Appraisal*, Wagner and Warner express a similar view:

> Participants in current discussions of naturalism seem to assume that the meaning of 'naturalism' ('naturalist program', etc.), its motivations and—often—its correctness, one way or the other, are almost obvious. The historical situation makes such assumptions exceedingly unlikely. Philosophers have taken just about every possible stance with some manner of justification, and all of the main programs within this area ("naturalism," "phenomenology," "analytic philosophy," and so forth) have been open to sharp differences of interpretation by their adherents. (Wagner & Warner, 1993, p. 3)

In a similar vein, David Papineau (1993) begins his book *Philosophical Naturalism* with the words, "What is philosophical 'naturalism'? The term is a familiar one nowadays, but there is little consensus on its meaning. . . . I suspect that the main reason for the terminological unclarity is that nearly everybody nowadays wants to be a 'naturalist,' but the aspirants to the term nevertheless disagree widely on substantial questions of philosophical doctrine" (p. 1).

These comments are amply verified by experience at philosophical meetings and institutes, and in reading the literature. Fred Dretske (1995), for example, takes the reductionist position that we understand something, including the mind, only when we can resolve it into its component parts and then see how their behavior generates the behavior of the whole. Ruth Millikan (1984), on the other hand, advocates a non-reductionist, non-mechanist, and externalist account of psychological categories in biological terms. She regards this, however, not as an alternative to naturalism, but as a variety of it. And Jaegwon Kim (1993) identifies the "natural" with the causal to the extent that a God who created the universe would thereby be a natural being, since creation is a kind of causation! And even John Searle's *Rediscovery of the Mind*—wherein he argues that there is something irreducibly first-person and experiential about consciousness that cannot be rendered in third-person discourse from the natural sciences—also (without further explanation) holds to the claim that the mind is nothing more than a causal product of the brain. So the initial impression that naturalism is almost a consensus view among philosophers of mind quickly gets more complicated upon closer inspection. Yes, these writers share a label, a respect for modern science, and a commitment to the vague notion that the mind must in *some sense* be "accommodated" within the bounds of "nature as conceived by the natural sciences." But it is not clear whether there is more commonality here than a commitment to a label, a kowtow, and a slogan. Or, less theatrically, if there is agreement that naturalism consists of a project of accommodating the mind within the world of nature as described by

natural science, there is no such agreement about (a) *what aspects of the mind need to be so accommodated* (e.g., consciousness, dispositional states like beliefs and desires, free will and action, occurrent states like sensations, perceptions and judgments, emotions, or personalities?), (b) *what would count as "accommodation"* (e.g., would this be some kind of explanation, some kind of metaphysical necessity, or some kind of methodology?), and (c) *what counts as the hallmarks of "the natural sciences"* (e.g., laws, causation, micro-explanation, or commitment to a materialistic ontology?) There are certainly a number of possible positions to be taken within the bounds of the naturalist slogan; and as Wagner and Warner (1993) point out, philosophers tend to take every available position they can find.

Papineau (1993) and Wagner and Warner (1993) also point out that conversations about naturalism often proceed as though both its meaning and its truth did not require prior discussion. Thus, critics of one sort of naturalism often find themselves at cross-purposes with naturalists of a different stripe. It is indeed possible that instead of having an actual consensus *view* here, it could turn out that for *every* naturalistic view proposed it is broadly rejected, not only by anti-naturalists, but also by naturalists of different stripes. That is, it could turn out that there is no common core to naturalistic views, but merely a name and a slogan. Indeed, one participant at the 1993 NEH Summer Institute on Naturalism, Jesse Hobbs, presented a paper entitled "Naturalism: A Contemporary Shibboleth?" The title ended with a question mark, but it captured nicely the observation of most participants that it was not at all clear whether there was a common view called "naturalism."

It is true, I think, that the word 'naturalism' does tend to function as a kind of shibboleth—that is, as a word whose use distinguishes "members of the tribe" from outsiders. And it is, I think, true that naturalism has become a kind of ideology in philosophical circles—that is, it is a widely-shared commitment to a way of believing, speaking, and acting whose basic assumptions are seldom examined or argued for. And a recent survey article on naturalism in philosophy of science describes Nagel's presidential address, cited above, as arguing against the objection that, "in committing itself to the logic of scientific proof without further foundations, naturalism is quite analogous to religious belief in resting on unsupported and undemonstrable faith" (Rosenberg, 1996, p. 1).

Naturalism Clarified

It is my view, however, that substantial order can be brought to this apparent mishmash. In a forthcoming book, *Mind and the World of Nature*, I tell a longer and more historical story about the history of contemporary naturalism (Horst, forthcoming). Here, however, I shall confine myself to the shape of the problem in the past several decades.

Naturalism, in rough characterization, is the thesis that *the mind can be accommodated within the framework of the world of nature as described by the natural sciences*. However, this first formulation is too broad. One might, after all, define 'nature' as Descartes did at one point, as something like "God and every created thing," or, as Jaegwon Kim (1993) has suggested, as the entire causal nexus. Such definitions would trivially include things as "natural" that the overwhelming majority of self-styled naturalists would want to reject as unnaturalistic: entities such as Cartesian souls, angels, God, or positions like idealism and pragmatism. So we must also add an additional caveat to rule out definitions of 'nature' that are so broad as to make naturalism trivial by allowing *super*natural entities. Thus, we shall add the caveat: A naturalist theory cannot be one that (a) posits the existence of supernatural entities, such as God, angels or immaterial souls, or (b) adopts a metaphysical stance in which the ontology of the natural sciences is not fundamental (e.g., transcendental idealism, pragmatism).

However, even within naturalism, our first characterization is not so definite as to be a solid philosophical *thesis*, but more of a thesis schema. For there are several axes along which this first characterization is yet unclear:

1. What is constitutive of "the framework of the natural sciences"?
2. What is meant by "accommodation"?
3. What kind of claim is it that the naturalist is making in uttering this schema? Is this a *positive* claim (about how the world in fact is), a *normative* claim, or a *methodological assumption*?

Three Paradigms of Science

If your aim is to accommodate the mind within the world described by the natural sciences, a lot is going to ride on how you conceive the natural sciences. In the 1930's through 50's, philosophers of science tended to speak in terms of a *unitary science*

that had a single set of objects, methods, and terms. This *unitary science* was not so much something that anyone thought they had in hand. Rather, it was a hoped for final state of science in which all of the "special" sciences, like chemistry, biology and psychology, would be shown to be special consequences of the *fundamental* science of basic physics. However, philosophers of science have increasingly come to recognize a diversity of methods and forms of explanation in different local sciences (e.g., biology versus physics), while history of science has revealed that those who we regard as the founding figures of particular sciences often had very different views of the nature of scientific explanation. There are perhaps five great watersheds in modern science: the developments of Galilean physics, Newtonian mechanics, Darwin's theory of evolution, Einstein's theory of relativity, and quantum mechanics. And each of these involved its own assumptions about the nature of scientific explanation. For example, Einstein rejected the idea that "God plays dice with the universe" in the fashion required by quantum randomness. The first three figures mentioned–Galileo, Newton, and Darwin–provide paradigms for science that form the basis of three different styles of naturalism, and I shall discuss these in order.[1]

Galileo is important to us not only for his particular scientific accomplishments but also for his advocacy of a method, sometimes called the "Method of Resolution and Composition." According to this methodology, we understand a thing when we can break it down into its component parts, understand the behavior of the parts, and then derive the behavior of the whole from the behavior of the parts. This view attempts to model the physical sciences upon geometry, in which one is able to construct and deduce complex mathematical objects and theorems from a simple set of definitions and axioms. This idea that explanation is fundamentally micro-explanation is clearly an idea that is reflected in the contemporary citations of Dretske (1995) and Fodor (1987) above, and indeed is the inspiration for a great variety of views that held sway for large parts of the twentieth century: logical behaviorism, psychofunctionalism, reductionism, type and token physicalism, and supervenience.

Yet, Galileo's method is not the only influential view of scientific explanation. In Newton, we find explicit disavowals of the need for, and even the desirability of hypothetical explanations appealing to unseen causes. For Newtonians, the goal of science is to provide mathematical laws that describe the regularities in observable phenomena–and to proceed no further. It is thus little surprise that the British Associationists of the eighteenth century, who were influenced primarily by Newton, were more interested in a "mental chemistry" or "mental geography" describing the inter-relations of mental states than in a reduction of the mind to the brain. The Newtonian model also re-emerges in the 19th-century Positivist philosophy of science of Mach and Comte. It later re-emerges in Watson and Skinner's behaviorist treatment of mental states as unobserved hypothetical entities that are unnecessary in scientific explanation, and looks only for regular connections between stimulus and response that are useful in the prediction and control of behavior. Arguably, this model also inspires those, like contemporary functionalists and computationalists, who view the mind in purely functional terms that are autonomous from physical form.[2]

A third paradigm of explanation is found in the evolutionary biology of Darwin, who produced explanations of things, not in terms of their structure, but in terms of their adaptation, selection, and reproductive history. To explain the sapsucker's bill, we do not look at the inner workings of the bill, but tell a story about how bills of a particular sort contribute to the chances of individual sapsuckers living long enough to reproduce (confers selective advantage), and then a story about how selective advantages conferred upon individuals in a population make a feature statistically likely to proliferate in future generations, as those members with the selec-

[1]I have nothing comparable to say about relativity or quantum mechanics. To the best of my knowledge, no one has drawn any general consequences for the mind from the theory of relativity. There are discussions of quantum theory and the mind–for example, some have argued that the apparent conflict between freedom and determinism is defused because contemporary physics does not view the physical world as deterministic, while others have claimed that the role of the observer in determining quantum events implies that the world is "mental" all the way down to the quantum level. However, these have not so much provided the basis for a breed of *naturalistic* theory, as the inspiration for some *anti*-naturalist arguments.

[2]Readers of the special issue may appreciate the oft-forgotten fact that Newton himself was a thorough going *super*naturalist: he did not believe, like Leibniz, that God wound the universe up like a giant clock and stood back to let it run. Rather, Newton believed that God had to continually intervene through countless acts of special providence to keep the planets in their courses! (cf. Dobbs & Jacob, 1995).

tive advantage are more successful at producing progeny. Although Darwin's own great contribution to evolutionary biology was that he provided a framework wherein species change and teleological categories could be grounded in concrete mechanisms, in practice evolutionary explanation is largely autonomous from reductions to an underlying mechanism. Evolutionary explanation was long practiced before the discovery of the DNA molecule as the basis for genetics, and even now we have at best a correlation between genes and phenotype, instead of an embryological story that *derives* the development of phenotypic features from DNA codes. This form of explanation was exploited early on in psychology by William James, and has recently received renewed attention in the writing of Ruth Millikan (1984), David Papineau (1993), Owen Flanagan (1992), and others.

The most central difference between these three paradigms lies in the fact that they involve three different standards of explanation: *reductive* (or "compositive," to use Galileo's term), *nomological* (i.e., based on laws), and *evolutionary*. The point is not that these three standards of explanation are incompatible with one another: for example, the laws of thermodynamics (Newtonian nomological paradigm) are sometimes held to be susceptible to a micro-explanation (Galilean paradigm) in terms of the statistical mechanics of gas molecule collisions. The point, rather, is that they present different views of what can count as an "explanation" in the natural sciences. And depending on which model you are thinking of, you may come to some very different conclusions about *what would be needed* in order to "accommodate" the mind within "the framework of the natural sciences." For the reductionist, to accommodate the mind in the framework of the natural sciences would be to show that mental phenomena are special complex cases of physical phenomena. For the Newtonian, it would merely be to bring mental phenomena under general laws–which is arguably no naturalization at all. For the evolutionary naturalist, it would be to show how mental phenomena arose as a result of a process of variation and selection.

Explanation and Metaphysics

The second unclarity about naturalism is about what is meant by "accommodating" the mind within nature. Naturalistic claims are sometimes advanced as claims about explanation, and sometimes as claims about metaphysics. *Metaphysical* claims are claims about how the world is. Naturalism as a metaphysical thesis is a claim that there is some sort of *determination relation* between facts cast in the vocabulary of the natural sciences and facts cast in mentalistic vocabulary: for example, that once you nail down the facts of basic physics, you have thereby nailed down all the facts of psychology as well. *Explanation*, by contrast, is a kind of *cognitive achievement* involving the production of insight in a human mind. Explanations come in many varieties, as they are answers to the various types of why-questions.

It is controversial whether there is a strong relationship between explanation and metaphysics. I happen to think that there *is* a strong relation between certain kinds of explanations that I call "conceptually adequate" and the metaphysical relation of metaphysical necessity. However, there are clearly types of explanation that do not entail any kind of metaphysical determination (e.g., statistical explanation). And it is at least conceivable that there are kinds of metaphysical determination that we are incapable of understanding or explaining. Hence, it is wise to separate these issues.

There is also an additional reason to treat them separately: in recent years, philosophers of mind have increasingly given grudging assent to the claim that there is an "explanatory gap" between mind and body – that things like conscious experience cannot be *explained* in naturalistic terms. (This issue will be discussed in the next section.) By and large, however, their response has been to retrench at the level of metaphysics, claiming that an explanatory gap does not entail the existence of a metaphysical gap between mind and body, but only a weakness of our understanding (McGinn, 1983; Nagel, 1974). Thus, it behooves us to separate questions about naturalistic *explanation* from questions about naturalistic *metaphysics*.

Positive and Normative Naturalism

The third point of clarification is about the overall *status* or *tenor* of the naturalist's claim. Naturalism can be taken as either a positive or a normative thesis. As a *positive* thesis, it is a claim *about how the world will actually turn out to be*: that we will, for example, be able–or would ideally be able, if only our minds did not have their current weaknesses–to explain the mental in terms of the physical. As

a *normative* thesis, it is the claim that the mental *must* be naturalized or something awful follows (e.g., that psychology cannot be a science, or that we are not entitled to our commitments to mental states, etc.).

I shall simply note in passing that *these two claims make poor bedfellows*: one cannot investigate naturalism as an empirical claim while also holding it as a normative thesis. The issue is this: *if push comes to shove between our commitment to naturalism and our commitment to, say, consciousness, which gives way?* If naturalism is a positive thesis (i.e., a claim about how things *are*), it is naturalism itself that stands in the dock as a metatheoretical thesis to be tested against the evidence, including the evidence from psychology. But if naturalism is a normative thesis, it is the (putative) evidence—including our commitment to mental states—that stands in the dock until they can be proven compatible with naturalistic metatheory. It seems to me that it is our broad metatheoretical speculations like naturalism that *ought* to stand in the dock against the evidence, and not vice-versa. Naturalists often present their claims as positive claims, but, as some of the quotes presented earlier attest, conformity with a naturalistic norm, such as reduction, or falling under strict laws, is often taken as a litmus for scientific and even ontological legitimacy. Because no *a priori* arguments are offered for this normative claim, I shall simply dismiss it, and concentrate on the positive claim here.

The Current State of Play in Philosophy of Mind: Reductive Explanation and Materialism

The most influential conversations in 20th century philosophy of mind were about the prospects for naturalizing the mind in the Galilean sense of reducing it to something else. On the side of explanation, this meant that the naturalist was committed to explaining features of the mind like language, consciousness, meaning, and intelligence in physical terms, probably by way of neuroscience or computational theory. On the metaphysical side, the naturalist was committed to the view that there is nothing in the world except physical stuff, and that as a consequence, mental states and processes are nothing over and above physical states and processes, albeit of a high degree of complexity.

I shall not belabor the internecine disputes between reductionists of different stripes. What is important for our purposes here is that, in the last decade or so of the 20th century, there was a growing admission, even among die-hard naturalists, that there is a problem in giving a naturalistic explanation for many features of the mind. While it seemed that neuroscience and information theory might be able to produce explanations of what the mind *does* in the sense of "how it reacts to environment," it was becoming increasingly clear that there was a real problem in trying to explain in physical terms what experience is like *from the inside*, from the first-person perspective. There are also problems about meaning and free will, but I shall concentrate here on the problem of conscious experience, which has received the lion's share of the attention.

This problem was most dramatically motivated in a series of now famous thought-experiments. The most important of these is most likely one proposed by Frank Jackson (1982). Jackson proposes that we imagine that there is a brilliant neuroscientist named Mary who lives sometime in the future. Mary knows (we stipulate) everything there is to be known about the nervous system—that is, she knows all the physical and neurological facts. However, Mary has been locked since birth in an environment that is carefully controlled so as to contain only the colors black and white and various shades of gray. She has thus never *seen* any chromatic colors herself, even though she understands the neurological processes involved in color vision. One day, her captivity is ended, and she is presented with a brilliant red object. She has a kind of visual experience she has never had before, and now she knows what it is like to see red. Now we pose the question: *has Mary learned something new here?* Namely, has she learned *what it is like to see red?* Or could she have *inferred* this from her vast neuroscientific knowledge? Jackson urges us that (a) she *did* learn something that she could not have reasoned on the basis of her knowledge of neuroscience (i.e., how red looks); (b) this shows that there are facts of subjective experience that are not explainable in terms of physical facts because if they were explainable in this way, then Mary could have derived the answer without actually experiencing red herself; and (c) this means that there are *facts* over and above the physical facts. (Note that (b) is a claim about explanation, whereas (c) is about metaphysics).

Thought-experiments of similar ilk were offered by other philosophers as well, such as Thomas Nagel

(1974), John Searle (1992), and David Chalmers (1996). Joseph Levine (1983) has canonized the problem as one of an "explanatory gap" (i.e., there is a gap between where explanations provided by the sciences of nature leave off, such as, at the level of neuroscience, and where facts about conscious experience begin). Moreover, these authors argue that the gap is not a mere accident of what we happen not to know at the moment, but rather it is *principled*: physical facts are not the right sorts of things to be even *potential* explainers of conscious experience. I shall condense a great deal of recent debate into a nutshell: *there has been increasing acceptance even among naturalists that there may well be a robust explanatory gap, and that micro-explanation of the mind in terms of the neurological activity of the brain does not seem suited to bridging this gap*.

Curiously, however, this has not caused many to abandon the naturalist camp. Instead, it has caused naturalists to seek other avenues to explore. One tack naturalists have taken is to emphasize that the gap that Jackson and others have pointed out is at the level of explanation rather than metaphysics. Some, such as Colin McGinn and Thomas Nagel, tend to locate the source of the gap in an incapacity of human understanding, and perhaps in a principled difficulty in any thinking system being able to grasp its own level of complexity. Such naturalists admit that conscious experience cannot be *explained*, but continue to hold to materialism as a *metaphysical* stance. They believe that the physical facts totally determine the facts about experience, but our minds are limited in ways that prevent us from seeing how this must be so. This position may well be consistent (i.e., it may not be self-contradictory), but it is important to see that the naturalist makes this move at the expense of making his materialism a kind of stance of faith, one that cannot be verified or falsified through empirical means.

A second approach is to seek some form of explanation *other than* or *in addition to* micro-reduction in terms of which one might naturalize the mind. This approach is in part motivated by recent work in philosophy of science. Reductive naturalism of the mind was in many ways motivated by reductionist views in the philosophy of science, which prevailed in the 1940's and 1950's. However, the past thirty years of philosophy of science have left reductionism pretty thoroughly discredited as a thesis about how even the natural sciences themselves proceed! If reduction is rare even at the level of chemistry and physics, and is outright contrary to the practice of biology and the life sciences, then it is particularly odd that philosophers of mind have held it up as a litmus for the legitimacy of psychology. It sometimes seemed as though the philosophy of mind of the 1990's was the last bastion of the philosophy of science of the 1950's.

The Newtonian Paradigm

One view of science that has barely outlasted reductionism is the Empiricist view, inspired by Newton, that the entirety of the scientific project consists in subsuming events under strict laws of nature. I personally think that if the notion of an empirical generalization is properly understood, this roughly Newtonian view of science still has a great deal to be said for it (see Horst, 1996, forthcoming). However, it should readily become clear that this view of science poses no threat of reducing the mind to something else. Suppose that we were to carry out a successful program of finding lawful relationships between one psychological state and another, or between psychological states and brain states. By "lawful relationships," we merely mean that when we have one, we also have the other. This kind of "lawful relationship," however, is precisely agnostic about the metaphysics of the relationship, in exactly the same way that Newton decided to be agnostic about the presence of attractive *forces* that explained the laws of planetary motion.

The spirit of this Newtonian program is to draw a sharp line between what is empirically testable (e.g., whether there is a regular relationship between *this* kind of activity in the cone cells in the eye and *that* kind of color experience) and what is not testable (e.g., the color experience just is a kind of neurological activity, a *non*-physical event is *caused* in a lawful way by physical events, etc.). This is a kind of demarcation between empirical science and philosophy. Elsewhere I have endorsed a version of this Newtonian interpretation of certain parts of psychology (cf. Horst, 1996, forthcoming). My point here is merely to point out that, if the naturalist cannot bridge the explanatory gap through reduction, he or she should look for no solace from the Newtonian view of science. This view of science does not even *attempt* to bridge explanatory gaps of this kind. Indeed, it distinguishes itself from reductionism by explicitly *countenancing* such gaps.

The Darwinian Paradigm

The other possible recourse for the naturalist is to seek an explanation of conscious experience through the third important explanatory paradigm: the Darwinian paradigm. Evolutionary explanation is an attempt to explain the differentiation of species over time (e.g., that these two populations of birds have different phenotypic traits). The Darwinian hypothesis is that terrestrial species stem from common ancestors, and hence evolutionary explanation seeks to render intelligible the diversity of species today against the assumption that they spring from a common stock. A phenotypic trait present in a species—say, a particular anatomical formation like a curved beak or a particular instinctive behavior—is viewed as the expression of a gene encoded in the DNA of particular organisms. Darwinian explanation posits that changes in species can be explained by two mechanisms that operate upon a population: variation and selection. Variation is a mechanism that produces random genetic changes, which are then expressed as differences in phenotypic traits. These differences can make the organisms in which the mutations occur more or less fit (i.e., able to survive long enough to pass on their genes) in their particular environment than those who do not have the mutation. As a result, selective forces operate at a statistical level to make it more likely that organisms that have a trait that makes them more fit will pass on their genes. The general schema of the explanation of a trait thus has three parts: (a) the trait is the expression of a particular gene (embryology and development), (b) this gene arose at some point in evolutionary history through a process of random mutation (variation), and (c) its expression conferred some advantage in fitness upon those organisms that possessed it, so that it proliferated in the population (selection).

I should note that the best of evolutionary theorists are generally more cautious than their reductionist cousins: they generally do not claim that *everything* about an organism—or indeed, even the survival of every genetic lineage—can be explained in this way. Some features are thought to persist, for example, because they are "free riders" on a genetic sequence that was selected for other benefits it provided (e.g., sickle cell anemia was not itself selected for, but is a by-product of a genetic sequence that provides immunity to malaria). And, of course, the process of variation requires that there be genes in the population that have not been selected for.

Evolutionary Explanation and Conscious Experience

What I wish to argue here is this: if *reductive* naturalism cannot explain conscious experience, it can expect no help from *evolutionary* explanation. The basic reason stems from the fact that the Darwinian explanation of a phenotypic feature requires three components: (a) a developmental mechanism for the expression of the gene, (b) a mechanism for variation that could produce it, and (c) a mechanism for selection. And it is really only (c)—the mechanisms of selection—that are really fully in the purview of the theory of evolution. The theory makes use of *assumptions* about development and variation, but these are really *promissory notes* to be filled in by more basic sciences like molecular genetics and embryology. A mechanism for selection cannot get off the ground if a gene cannot arise through variation and be expressed through development in the first place. *But variation and development are precisely the sorts of events that call for broadly reductive explanations*; and so if there is an explanatory gap between physics (or biology or neuroscience) and conscious experience, this means in part that conscious experience cannot be explained by genetics and development, and so the explanation cannot proceed to the stage in which the factors truly proper to the theory of evolution (i.e., mechanisms of selection) would come into play.

Let us consider the matter a little more concretely. How would one supply an evolutionary explanation of something like consciousness? One would do so by singling out some selective advantage that consciousness confers on those that have it as compared to those that lack it, and then hypothesize that: (a) at some stage of our evolution, genes for consciousness appeared through a process of random variation (perhaps suddenly, perhaps through a long series of small changes); (b) those that possessed this gene proved, at a statistical level, to be more fit than those that lacked it in the environment in which our ancestors found themselves during this crucial period; and hence (c) the gene that controls this trait proliferated in the human phenotype (cf. Dretske 1995; Flanagan 1992). The *explanation* here is of the following form: "Given assumptions a and b, we can explain c." The implicit rationale through which the plausibility of the argument is generally filled out goes

something like this: "We know that c is the case, and we need to explain it. We just see how reasonable b is—it is almost a truth of reason that such a trait would confer such-and-such a selective advantage. And this trait must have emerged *somehow* and at *some point*. Therefore, it's reasonable to postulate a. And as a result, we have a plausible explanation."

It is by now well known how easy it is to abuse this explanation form to generate too easy "just so stories," particularly those rampant in evolutionary psychology (Kitcher, 1985). What I wish to point out, however, is something completely different: namely that this kind of explanation is no more plausible than is premise a—the premise that some process of genetic variation, combined with a process of development, *could* produce the feature in question, in this case, consciousness. That is, the evolutionary explanation depends upon the availability-in-principle of an explanation appealing to molecular genetics and embryology. It does not require that one have the explanation on-hand, of course,—evolutionary explanations almost always outstrip our knowledge of embryology and molecular genetics—but it does require that such an explanation be available in principle. If *no possible biochemical event* could produce a gene whose expression could account for the consciousness of an organism, then step a of the explanation cannot get off the ground, and hence we never get to the point where the mechanism for selection can come into play.

An example will make this clear and even memorable. Suppose there were a gene that endowed its possessors with an energy source in the form of a perpetual motion machine. Clearly, such a feature would be very valuable, and would confer selective advantage: organisms that had it would not need to eat or perhaps even to rest, and so would have much more time and attention to give to the crucial evolutionary activities of finding mates and avoiding predators. Yet, we would be skeptical of an explanation that appealed to such a gene—*not* because of problems with how it would come into play in natural selection, but because we have principled reason to doubt that any physical system could give rise to a perpetual motion machine in the first place, and so none could muster the entry conditions for the process of natural selection.

I suggest that the situation is very similar when we postulate a gene for consciousness. Or, more carefully, the situations are very similar *if we cannot cross the explanatory gap between the physical world and conscious experience*. For if there is a principled gap here, then there *is* principled reason to think that there is no gene and no developmental process that *could* account for the appearance of consciousness as a trait in even a single organism. This is just the sort of thing that is at stake in an explanatory gap between mind and matter. Evolutionary explanation could explain the *proliferation* of consciousness in a population *if* its first appearance could be explained at a more basic and reductive level. But if there is a robust explanatory gap, then the appearance of consciousness cannot be explained at a more basic and reductive level, and hence evolutionary explanation never gets off the ground.

The problem I have identified here (see also Horst, 1999) is not one that appears if you look *only* at the truly *evolutionary* side of evolutionary explanation (i.e., the selective side). The problem appears when you examine the preconditions that have to be met in order for selection to get off the ground in the first place (i.e., the assumptions that there are physical mechanisms that could produce genes whose expression would result in the features that we wish to explain, in this case conscious experience). Here, too, one might have been tempted to leave large promissory notes to be filled in later, as is generally the case in evolutionary explanation. However, it is here that evolutionary naturalism runs into the problems that have already become manifest for reductive naturalism: the problem of the explanatory gap. For the promissory note left by the evolutionary naturalist is one that the reductive naturalist would have to make good on, and it is precisely here that the reductive naturalist has begun to doubt that he possesses proper currency to pay such a debt. Thus, *if* it is in fact the case that reductive explanation cannot explain conscious experience, the naturalist cannot look to evolutionary explanation for help here, because the problems are precisely in the areas in which the evolutionary naturalist would defer to a reductionist program to be specified at a later date.

Conclusion

Let me try briefly to summarize what I've tried to accomplish in this paper. First, I've tried to clarify different strands of naturalism (i.e., reductive, nomological, Darwinian) in contemporary philosophy of

mind against an historical backdrop. Second, I've described the current "state of play" in the field in which it is widely admitted that there seems to be an "explanatory gap" that makes it impossible to explain conscious experience in terms of physical or neurological phenomena. I happen to think that this explanatory gap is a robust one that we will not be able to find a way beyond. However, I have not argued for that here, but rather for a more modest conclusion: namely, third, that *if* there is a robust explanatory gap that does not allow for (broadly) reductive explanations of conscious experience, this gap cannot be filled in by evolutionary explanation. Why not? Because evolutionary explanation only gets off the ground when we are dealing with phenotypic features whose appearance and expression could, in principle, be explained in (broadly) reductive terms. But this is just what is denied if there is an irreducible explanatory gap.

References

Chalmers, D. (1996). *The Conscious Mind*. New York: Oxford University Press.

Churchland, P. (1981). Eliminative materialism and propositional attitudes. *Journal of Philosophy, 78*, 67-90.

Dobbs, B. J. T., & Jacob, M. C. (1995). *Newton and the culture of Newtonianism*. Atlantic Highlands, NJ: Humanities Press.

Dretske, F. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Flanagan, O. (1992). *Consciousness reconsidered*. Cambridge, MA: MIT Press.

Fodor, J. (1987). *Psychosemantics*. Cambridge, MA: Bradford Books.

Horst, S. (1996). *Symbols, computation and intentionality: A critique of the computational theory of mind*. Berkeley, CA: The University of California Press.

Horst, S. (1999). Evolutionary explanation and the hard problem of consciousness. *Journal of Consciousness Studies, 6* (1), 38-48.

Horst, S. (forthcoming). *Mind and the world of nature* [On-line]. Manuscript in preparation. Available: http://www.wesleyan.edu/~shorst

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly, 32*, 127-136.

Kim, J. (1993). Supervenience and mind: Selected philosophical essays. New York: Cambridge University Press.

Kitcher, P. (1985). *Vaulting ambition: Sociobiology and the quest for human nature*. Cambridge, MA: MIT Press.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly, 64*, 354-61.

McGinn, C. (1983). *The subjective view*. New York: Clarendon Press.

Millikan, R. (1984). *Language, thought and other biological categories*. Cambridge, MA: MIT Press.

Nagel, E. (1956). Presidential address. *Proceedings of the American Philosophical Association,* 28, pp. 5-17.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435-450.

Papineau, D. (1993). *Philosophical naturalism*. New York: Blackwell.

Rosenberg, A. (1996). A field guide to recent species of naturalism. *British Journal for the Philosophy of Science, 47* (1), 1-29.

Searle, J. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Stich, S., & Laurence, S. (1994). Intentionality and naturalism. In P. A. French, T. E. Uehling, Jr., & H. K. Wettstein (Eds.), *Midwest studies in philosophy volume XIX: Philosophical Naturalism* (pp. 159-182). Notre Dame, IN: University of Notre Dame Press.

Wagner, S., & Warner, R. (Eds.). (1993). *Naturalism: A critical appraisal*. Notre Dame, IN: University of Notre Dame Press.

Author

HORST, STEVEN. Address: Department of Philosophy, Wesleyan University, Middletown, CT 06459.
Title: Associate Professor of Philosophy.
Degrees: Ph.D., University of Notre Dame.
Specializations: Pholosopy of mind, philosophy of science, cognitive science, metaphysics.