# Chapter 5 — Evolution versus Naturalism

Alvin C. Plantinga

5

EVOLUTION VERSUS NATURALISM[1]

ALVIN C. PLANTINGA

Take *philosophical naturalism* to be the belief that there is no such person as God or anything like God. Naturalism, so taken, is stronger than atheism; it is possible to be an atheist without rising to the heights (or sinking to the depths) of naturalism. A follower of Hegel could be an atheist, but because of his belief in the Absolute would not qualify for naturalism; similarly for someone who believed in the Stoic's *Nous,* or Plato's Idea of the Good or Aristotle's Prime Mover who thinks only about himself. This definition of naturalism, of course, is a bit vague: how much like God must an entity be so that endorsing it disqualifies one from naturalism? Perhaps the definition will be serviceable nonetheless; clear examples of naturalists would be the Bertrand Russell of "A Free Man's Worship," Daniel Dennett (*Darwin's Dangerous Idea*), Richard Dawkins (*The Blind Watchmaker*), the late Stephen Jay Gould, David Armstrong, and the many others that are sometimes said to endorse "The Scientific Worldview."[2]

Now naturalism is not, as it stands, a religion. Nevertheless it performs one of the most important functions of a religion: it provides its adherents with a worldview. It tells us what the world is fundamentally like, what is most important and deep in the world, how we are related to other creatures in the world, what our place in the world is, what (if anything) we can expect after death, and the like. A religion typically does that and more; it also involves worship and ritual. These latter are ordinarily (but not always) absent from naturalism; naturalism, we could therefore say, performs the cognitive functions of a religion. Naturalism has a creed, nicely stated by Peter van Inwagen (himself no naturalist):

There is no God. There is, in fact, nothing besides the physical cosmos that science investigates. Human beings, since they are a part of this cosmos, are physical things and therefore do not survive death.… Like other animals, they are a product of uncaring and unconscious physical processes that did not have them, or anything else, in mind.… Human beings, however, have an unfortunate tendency to wish to deny these facts and to believe comforting myths according to which they have an eternal purpose. This irrational component in the psyches of most human beings—it is the great good fortune of the species that there are a few strong-minded progressives who can see through the comforting myths—encourages the confidence game called religion.… [3]

Since naturalism performs the cognitive or worldview function of religion, we can perhaps promote it to the status of an honorary religion. What I propose to argue is that naturalism and contemporary evolutionary theory are at serious odds with one another—and this despite the fact that the latter is ordinarily thought to be one of the main supporting members in the edifice of the former.[4] Accordingly, there is a religion/science conflict in the neighborhood of evolution—not, however, a conflict between evolution and Christian belief, but between naturalism and evolution. (Of course I am *not* attacking the theory of evolution, or the claim that human beings have evolved from simian ancestors, or anything in that neighborhood; I am instead attacking the conjunction of *naturalism* with the view that human beings have evolved in that way. I see no similar problems with the conjunction of *theism* and the idea that human beings have evolved in the way contemporary evolutionary science suggests.)

More particularly, I'll argue that the conjunction of naturalism with the belief that human beings have evolved in conformity with current evolutionary doctrine—"evolution" for short—is in a certain interesting way *self-defeating* or self-referentially incoherent. Still more particularly, I'll argue that naturalism and evolution—"N&E" for short—furnishes one who accepts it with a *defeater* for the belief that our cognitive faculties are reliable—a defeater that can't be defeated. But then this conjunction also furnishes a defeater for any belief produced by our cognitive faculties, including, in the case of one who accepts it, N&E itself: hence its self-defeating character.

## 1. P(R/N&E) IS LOW

So much for a quick overview of the argument. More specifically EAAN begins from certain doubts about the *reliability* of our cognitive faculties, where, roughly,[5] a cognitive faculty—memory, perception, reason—is reliable if the great bulk of its deliverances are true. These doubts are connected with the *origin* of our cognitive faculties. According to current evolutionary theory, we human beings, like other forms of life, have developed from aboriginal unicellular life by way of such mechanisms as natural selection and genetic drift working on sources of genetic variation: the most popular is random genetic mutation. Natural selection discards most of these mutations (they prove deleterious to the organism in which they appear), but some of the remainder turn out to have survival value and to enhance fitness; they spread through the population and thus persist. As this story has it, it is by way of these mechanisms, or mechanisms very much like them, that all the vast variety of contemporary organic life has developed; and it is by way of these same mechanisms that our cognitive faculties have arisen.

Now according to traditional Christian (and Jewish and perhaps Muslim) thought, we human beings have been created in the image of God. This means, among other things, that he created us with the capacity for achieving *knowledge*—knowledge of our environment by way of perception, of other people by way of something like what Thomas Reid calls *sympathy*, of the past by memory and testimony, of mathematics and logic by reason, of morality, our own mental life, God himself, and much more.[6] And of course most of us are inclined to think that our cognitive faculties, our belief-producing processes, are for the most part reliable. True, they may not be reliable at the upper limits of our powers, as in some of the more speculative areas of physics; and the proper function of our faculties can be skewed by envy, hate, lust, mother love, greed, and so on. But over a broad area of their operation, we think the purpose of our cognitive faculties is to furnish us with true beliefs, and that when they function properly, they do exactly that.

The above evolutionary account of our origins, clearly enough, is compatible with the theistic view that God has created us in his image.[7] So evolutionary theory taken by itself (without the patina of philosophical naturalism that often accompanies expositions of it) is not as such in tension with the idea that God has created us and our cognitive faculties in such a way that the latter are reliable, that (as the medievals liked to say) there is an adequation of intellect to reality. But if *naturalism* is true, there is no God, and hence no God (or anyone else) overseeing our development and orchestrating the course of our evolution. And this leads directly to the question whether it is at all likely that our cognitive faculties, given naturalism and given their evolutionary origin, would have developed in such a way as to be reliable, to furnish us with mostly true beliefs. The problem, from this point of view, is that the ultimate purpose or function of our cognitive faculties, if they have one, is not to produce true beliefs, but to promote reproductive fitness. As evolutionary psychologist David Sloan Wilson puts it, “the well-adapted mind is ultimately an organ of survival and reproduction.”[8] What our minds are *for* is not the production of true beliefs, but the production of adaptive behavior. That our species has survived and evolved at most guarantees that our behavior is adaptive; it does not guarantee or even suggest that our belief-producing processes are for the most part reliable, or that our beliefs are for the most part true. That is because, obviously, our behavior could be adaptive, but our beliefs mainly false. Darwin himself expressed this doubt: “With me,” he said, “the horrid doubt always arises whether the convictions of man’s mind, which has been developed from the mind of the lower animals, are of any value or at all trustworthy. Would any one trust in the convictions of a monkey’s mind, if there are any convictions in such a mind?”[9]

The same thought is put more explicitly by Patricia Churchland. She insists that the most important thing about the human brain is that it has evolved; this means, she says, that its principal function is to enable the organism to *move* appropriately:

Boiled down to essentials, a nervous system enables the organism to succeed in the four F's: feeding, fleeing, fighting and reproducing. The principle chore of nervous systems is to get the body parts where they should be in order that the organism may survive.… Improvements in sensorimotor control confer an evolutionary advantage: a fancier style of representing is advantageous *so long as it is geared to the organism's way of life and enhances the organism's chances of survival* [Churchland's emphasis]. Truth, whatever that is, definitely takes the hindmost.[10]

What Churchland means, I think, is that evolution is directly interested (so to speak) only in adaptive *behavior* (in a broad sense including physical functioning) not in true belief. Natural selection doesn't care what you *believe;* it is interested only in how you *behave*. It selects for certain kinds of behavior: those that enhance fitness, which is a measure of the chances that one's genes are widely represented in the next and subsequent generations. It doesn't select for belief, except insofar as the latter is appropriately related to behavior. But then the fact that we have evolved guarantees at most that we *behave* in certain ways—ways that contribute to our (or our ancestors') surviving and reproducing in the environment in which we have developed. Churchland's claim, I think, can perhaps be understood as the suggestion that the objective[11] probability that our cognitive faculties are reliable, given naturalism and given that we have been cobbled together by the processes to which contemporary evolutionary theory calls our attention, is low. Of course she doesn't explicitly mention naturalism, but it certainly seems that she is taking it for granted. For if theism were true, God might be directing and orchestrating the process in such a way as to produce, in the long run, beings created in his image and thus capable of knowledge; but then it wouldn't be the case that truth takes the hindmost.

We can put Churchland's claim as

P(R/N&E) IS LOW,

where "R" is the proposition that our cognitive faculties are reliable, "N" the proposition that naturalism is true, and "E" the proposition that we have evolved according to the suggestions of contemporary evolutionary theory. I believe this thought—the thought that P(R/N&E) is low—is also what worries Darwin in the above quotation: I shall therefore call it "Darwin's Doubt."

Are Darwin and Churchland right? Is this probability low? Well, they are certainly right in thinking that natural selection is directly interested only in behavior, not belief, and that it is interested in belief, if at all, only indirectly, by virtue of the relation between behavior and belief. If adaptive behavior guarantees or makes probable reliable faculties, then perhaps P(R/N&E) will be rather high: we, or rather our ancestors, engaged in at least reasonably adaptive behavior, so it must be that our cognitive faculties are at least reasonably reliable, in which case it is likely that most of our beliefs are true—at least for those beliefs connected with survival and reproduction. (This likelihood, therefore, doesn't extend to philosophical beliefs, or highly abstract scientific or mathematical beliefs; it is only the occasional assistant professor of logic who needs to know how to prove, say, Gödel's theorems in order to survive and reproduce.) On the other hand, if our having reliable faculties *isn't* guaranteed by or even particularly probable with respect to adaptive behavior, then presumably P(R/N&E) will be rather low. If, for example, behavior isn't caused or governed by belief, the latter would be, so to speak, invisible to natural selection; in that case it would be unlikely that the great preponderance of true belief over false required by reliability would be forthcoming.[12] So the question of the value of P(R/N&E) really turns on the relationship between belief and behavior. Our having evolved and survived makes it likely that our cognitive faculties are reliable and our beliefs are for the most part true, only if it would be impossible or unlikely that creatures more or less like us should behave in fitness-enhancing ways but nonetheless hold mostly false beliefs.[13]

*Is* this impossible or unlikely? That depends upon the relation between belief and behavior. What would or could that relation be? To try to guard against interspecific chauvinism, I suggest that we think, not about ourselves and our behavior, but about a population of creatures a lot like us on a planet a lot like Earth (Darwin suggested we think about monkeys in this connection). These creatures are *rational:* that is, they form beliefs, reason, change beliefs, and the like. We imagine furthermore that they and their cognitive systems have evolved by way of the mechanisms to which contemporary evolutionary theory directs our attention, unguided by the hand of God or anyone else. Now what is P(R/N&E), specified not to us, but to them? To answer, we must think about the relationship between their beliefs and their behavior. There are four mutually exclusive and jointly exhaustive possibilities.

(1) One possibility is *epiphenomenalism:*[14] their behavior is not caused by their beliefs. On this possibility, their movement and behavior would be caused by something or other—perhaps neural impulses—which would be caused by other organic conditions including sensory stimulation: but belief would not have a place in this causal chain leading to behavior. This view of the relation between behavior and belief (and other mental phenomena such as feeling, sensation, and desire) is currently rather popular, especially among those strongly influenced by biological science. *Time* (December 1992) reports that J. M. Smith, a well-known biologist, wrote "that he had never understood why organisms have feelings. After all, orthodox biologists believe that behavior, however complex, is governed entirely by biochemistry and that the attendant sensations —fear, pain, wonder, love—are just shadows cast by that biochemistry, not themselves vital to the organism's behavior.…" He could have added that (according to biological orthodoxy) the same goes for beliefs—at least if beliefs are not themselves just biochemical phenomena. If this way of

thinking is right with respect to our hypothetical creatures, their beliefs would be *invisible* to evolution; and then the fact that their belief-forming mechanisms arose during their evolutionary history would confer little or no probability on the idea that their beliefs are mostly true, or mostly nearly true. Indeed, the probability of those beliefs being for the most part true would have to be rated fairly low (or inscrutable). On N&E and this first possibility, therefore, the probability of R will be rather low.

(2) A second possibility is *semantic* epiphenomenalism: it could be that their beliefs do indeed have causal efficacy with respect to behavior, but not by virtue of their *content.* Note first that nearly all naturalists are also *materialists* about human persons; we human beings are not immaterial Cartesian substances, and do not contain an immaterial part—a soul, for example. Here I'll assimilate materialism (about human beings) to naturalism; since all or nearly all naturalists are materialists, there will be little if any loss of generality. And from this point of view, a belief would have to be something like a long-term event or structure in the nervous system—perhaps a group of neurons connected and related in a certain way. This neural structure will have *neurophysiological* properties ("NP properties" for short): properties specifying the number of neurons involved in the structure, the way in which the neurons involved are connected with each other and with other structures (with muscles, sense organs, other neuronal events, etc.), the average rate and intensity of neuronal firing in various parts of this event and the ways in which this changes over time and in response to input from other areas. Now of course it is easy to see how *these* properties of this neuronal event should have causal influence on the behavior of the organism. A given belief is neurally connected both with other beliefs and with muscles; we can see how electrical impulses coming from the belief can negotiate the usual neuronal channels and ultimately cause muscular contraction and behavior.

But if this belief is really a *belief,* then it will also have *other* properties, properties in addition to its neurophysiological properties. In particular, it will have *content;* it will be the belief that *p*, for some proposition *p*—perhaps the proposition *naturalism is all the rage these days.* Having such and such a content would be a *semantic* property of the belief. Beliefs, therefore, have two very different kinds of properties: neurophysiological properties on the one hand, and semantic properties on the other. And now the question is this: does a belief—a neural structure—cause behavior, that is, enter into the causal chain leading to behavior, *by virtue of its content?* According to semantic epiphenomenalism, the answer is no. Belief is indeed causally efficacious with respect to behavior, but by virtue of the *syntactic* properties of a belief, not its semantic properties. If the first possibility is widely popular among those influenced by biological science, this possibility is widely popular among contemporary philosophers of mind; indeed, Robert Cummins goes so far as to call it the "received view."[15]

On this second possibility, as on the first, P(R/N&E) (specified to those creatures) will be low. The reason is that truth and falsehood, of course, are among the *semantic* properties of a belief, not its syntactic properties. But if the former aren't involved in the causal chain leading to behavior, then once more beliefs—or rather, their semantic properties, including truth and falsehood—will be invisible to natural selection. But then it will be unlikely that their beliefs are mostly true and hence unlikely that their cognitive faculties are reliable. The probability of R on N&E together with this possibility (as with the last), therefore, will be relatively low.

(3) It could be that beliefs are causally efficacious—by virtue of content as well as neurophysiological properties—but *maladaptive:* from the point of view of fitness these creatures would be better off without them. The probability of R on N&E together with this possibility, as with the last two, would also seem to be relatively low.

(4) Finally, it could be that the beliefs of our hypothetical creatures are indeed both causally connected with their behavior and also adaptive. (I suppose this is the commonsense view of the connection between behavior and belief in our own case.) What is the probability (on this assumption—call it C—together with N&E) that their cognitive faculties are reliable; and what is the probability that a belief produced by those faculties will be true? In WPF I argued that this probability isn't nearly as high as one is initially inclined to think. The reason is that if behavior is caused by *belief*, it is also caused by *desire* (and other factors—suspicion, doubt, approval and disapproval, fear—that we can here ignore). For any given adaptive action, there will be many belief-desire combinations that could produce that action; and very many of those belief-desire combinations will be such that the belief involved is false.

So suppose Paul is a prehistoric hominid; a hungry tiger approaches. Fleeing is perhaps the most appropriate behavior: I pointed out that this behavior could be produced by a large number of different belief-desire pairs. To quote myself:

Perhaps Paul very much *likes* the idea of being eaten, but when he sees a tiger, always runs off looking for a better prospect, because he thinks it unlikely that the tiger he sees will eat him. This will get his body parts in the right place so far as survival is concerned, without involving much by way of true belief..… Or perhaps he thinks the tiger is a large, friendly, cuddly pussycat and wants to pet it; but he also believes that the best way to pet it is to run away from it … or perhaps he thinks the tiger is a regularly recurring illusion, and, hoping to keep his weight down, has formed the resolution to run a mile at top speed whenever presented with such an illusion; or perhaps he thinks he is about to take part in a 1600 meter race, wants to win, and believes the appearance of the tiger is the starting signal; or perhaps.… Clearly there are any number of belief-cum-desire systems that equally fit a given bit of behavior. (WPF 225–26)

Accordingly, there are many belief-desire combinations that will lead to the adaptive action; in many of these combinations, the beliefs are false. Without further knowledge of these creatures, therefore, we could hardly estimate the probability of R on N&E and this final possibility as high.

A problem with the argument as thus presented is this: It is easy to see, for just *one* of Paul's actions, that there are many different belief-desire combinations that yield it; it is less easy to see how it could be that most of all of his beliefs could be false but nonetheless adaptive or fitness enhancing. Could Paul's beliefs really be mainly false, but still lead to adaptive action? Yes indeed; perhaps the simplest way to see how is by thinking of systematic ways in which his beliefs could be false but still adaptive. Perhaps Paul is a sort of early Leibnizian and thinks everything is conscious (and suppose that is false); furthermore, his ways of referring to things all involve definite descriptions that entail consciousness, so that all of his beliefs are of the form *That so-and-so conscious being is such-and-such.* Perhaps he is an animist and thinks everything is alive. Perhaps he thinks all the plants and animals in his vicinity are witches, and his ways of referring to them all involve definite descriptions entailing witchhood. But this would be entirely compatible with his belief's being adaptive; so it is clear, I think, that there would be many ways in which Paul's beliefs could be for the most part false, but adaptive nonetheless. Religious belief is nearly universal across the world; according to naturalists it is false, but nonetheless adaptive. Our question is really about the proportion of true beliefs among adaptive beliefs. For every true adaptive belief, however, it seems that one can easily think of a false belief adaptive in the same circumstances. The fact that my behavior (or that of my ancestors) has been adaptive, therefore, is at best a third-rate reason for thinking my beliefs mostly true and my cognitive faculties reliable—and that is true even given the commonsense view of the relation of belief to behavior. So we can't sensibly argue from the fact that our behavior (or that of our ancestors) has been adaptive to the conclusion that our beliefs are mostly true and our cognitive faculties reliable. It is therefore hard to see that P(R/N&E&C) is very high. To concede as much as possible to the opposition, however, let's say that this probability is either inscrutable or in the neighborhood of 0.9.

What we have seen so far is that there are four mutually exclusive and jointly exhaustive possibilities with respect to that hypothetical population: epiphenomenalism simpliciter, semantic epiphenomenalism, the possibility that their beliefs are causally efficacious with respect to their behavior but maladaptive, and the possibility that their beliefs are both causally efficacious with respect to behavior and adaptive. P(R/N&E) will be the weighted average of $P(R/N\&E\&P_i)$ for each of the four possibilities $P_i$—weighted by the probabilities, on N&E, of those possibilities. The probability calculus gives us a formula here:

$$(1)\quad P(R/N\&E) = (P(R/N\&E\&P_1) \times P(P_1/N\&E)) + (P(R/N\&E\&P_2) \times P(P_2/N\&E)) + (P(R/N\&E\&P_3) \times P(P_3/N\&E)) + (P(R/N\&E\&P_4) \times P(P_4/N\&E)).$$

Of course the very idea of a calculation (suggesting, as it does, the assignment of specific real numbers to these various probabilities) is laughable: the best we can do are vague estimates. Still, that will suffice for the argument. Now let's agree that $P_3$—the proposition that belief enters the causal chain leading to behavior both by virtue of neurophysiological properties and by virtue of semantic properties, but is nevertheless maladaptive—is very unlikely; it is then clear from (1) that its contribution to P(R/N&E) can safely be ignored. Note further that epiphenomenalism and semantic epiphenomenalism unite in declaring or implying that the *content* of belief lacks causal efficacy with respect to behavior; the content of belief does not get involved in the causal chain leading to behavior. So we can reduce these two possibilities to one: the possibility that the content of belief has no causal efficacy. Call this possibility "-C." What we have so far seen is that the probability of R on N&E&-C is low or inscrutable, and that the probability of R on N&E&C is also inscrutable or at best moderately high. We can therefore simplify (1) to

$$(2)\quad P(R/N\&E) = [P(R/N\&E\&C) \times P(C/N\&E)] + [P(R/N\&E\&\text{-}C) \times P(\text{-}C/N\&E)],$$

i.e., the probability of R on N&E is the weighted average of the probabilities of R on N&E&C and N&E&-C—weighted by the probabilities of C and -C on N&E.

We have already noted that the left-hand term of the first of the two products on the right side of the equality is either moderately high or inscrutable; the second is either low or inscrutable. What remains is to evaluate the weights, the right-hand terms of the two products. So what is the probability of -C, given N&E; what is the probability of semantic epiphenomenalism on N&E? Note that, according to Robert Cummins, semantic epiphenomenalism is in fact the received view as to the relation between belief and behavior.[16] That is because it is extremely hard to envisage a way, given materialism, in which the content of a belief *could* get causally involved in behavior. According to materialism, a belief just is a neural structure of some kind—a structure that somehow possesses content. How can its content get involved in the causal chain leading to behavior? Had a given such structure had a *different* content, one thinks, its causal contribution to behavior would be the same. What causes the muscular contractions involved in behavior are physiological states of the nervous system, including physiological properties of those structures that constitute beliefs; the content of those beliefs appears to be causally irrelevant. It is therefore exceedingly hard to see how semantic epiphenomenalism can be avoided, given N&E. (There have been some valiant efforts,[17] but things don't look hopeful.) So it looks as if P(-C/N&E) will have to be estimated as relatively high; let's say (for definiteness) 0.7, in which case P(C/N&E) will be 0.3. Of course we could easily be wrong; we don't really have a solid way of telling; so perhaps the conservative position here is that this probability too is inscrutable: one simply can't tell what it is. Given current knowledge, therefore, P(-C/N&E) is either high or inscrutable. And if P(-C/N&E) is inscrutable, then the same goes, naturally enough, for P(C/N&E). What does that mean for the sum of these two products, i.e., P(R/N&E)?

Well, we have several possibilities. Suppose we think first about the matter from the point of view of someone who doesn't find any of the probabilities involved inscrutable. Then P(C/N&E) will be in the neighborhood of 0.3, P(-C/N&E) in the neighborhood of 0.7, and P(R/N&E&-C) perhaps in the neighborhood of 0.2. This leaves P(R/N&E&C), the probability that R is true given ordinary naturalism together with the commonsense or folk-theoretical view as to the relation between belief and behavior. Given that this probability is not inscrutable, let's say that it is in the neighborhood of 0.9. And given these estimates, P(R/N&E) will be in the neighborhood of 0.41.[18] Suppose, on the other hand, we think the probabilities involved are inscrutable: then we will have to say the same for P(R/N&E). P(R/N&E), therefore, is either low—less than .5, at any rate—or inscrutable.

## 2. A DEFEATER FOR R

Returning to Darwin's doubt, observe that if this is the sensible attitude to take to P(R/N&E) specified to that hypothetical population, then it will also be the sensible attitude towards P(R/N&E) specified to us. We are relevantly like them in that *our* cognitive faculties have the same kind of origin and provenance as *theirs* are hypothesized to have. And the next step in the argument is to point out that each of these attitudes—the view that P(R/N&E) is low and the view that this probability is inscrutable—gives the naturalist-evolutionist a *defeater* for R. It gives him a reason to doubt it, a reason not to affirm it. I argued this by analogy. Among the crucially important facts, with respect to the question of the reliability of a group of cognitive faculties, are facts about their *origin.* Suppose I believe that I have been created by an evil Cartesian demon who takes delight in fashioning creatures who have mainly false beliefs (but think of themselves as paradigms of cognitive excellence): then I have a defeater for my natural belief that my faculties are reliable. Turn instead to the contemporary version of this scenario, and suppose I come to believe that I have been captured by Alpha-Centaurian super-scientists who have made me the subject of a cognitive experiment in which the subject is given mostly false beliefs: then, again, I have a defeater for R. But to have a defeater for R it isn't necessary that I believe that in fact I *have* been created by a Cartesian demon or been captured by those Alpha-Centaurian superscientists. It suffices for me to have such a defeater if I have considered those scenarios, and the probability that one of those scenarios is true, is inscrutable for me—if I can't make any estimate of it, or do not have an opinion as to what that probability is. It suffices if I have considered those scenarios, and *for all I know or believe* one of them is true. In these cases too I have a reason for doubting, a reason for withholding[19] my natural belief that my cognitive faculties are in fact reliable.

Perhaps it seems harder to see that one has a defeater for R in the case where the relevant probability is inscrutable, than in the case where it is low. But suppose you buy a thermometer; then you learn that the owner of the factory where it was constructed is a Luddite who aims to do what he can to disrupt contemporary technology, and he therefore makes at least some instruments that are unreliable. You can't say what the probability is of this thermometer's being reliable, given that it was made in that factory; that probability is inscrutable for you. But would you trust the thermometer? It's outside your window, and reads 30 degrees F; if you have no other source of information about the temperature outside, would you believe it is 30 degrees F?

One more analogy: you embark on a voyage of space exploration and land on a planet revolving about a distant sun, a planet that seems to have a favorable atmosphere. You crack the hatch, step out, and immediately find what appears to be an instrument that looks a lot like a terrestrial radio; you twiddle the dials, and after a couple of squawks it begins to emit strings of sounds that, oddly enough, form English sentences. These sentences express propositions only about topics of which you have no knowledge: what the weather is like in Beijing at the moment, whether Caesar had eggs on toast on the morning he crossed the Rubicon, whether the first human being to cross the Bering Strait was left-handed, and the like. Unduly impressed, indeed awed, by your find, you initially form the opinion that this instrument speaks the truth, that the propositions expressed (in English) by those sentences are true. But then you recall that you have no idea at all as to who or what constructed the instrument, what it is for, whether it has a purpose at all. You see that the probability of its being reliable, given what you know about it, is inscrutable. Then you have a defeater for your initial belief that the thing does in fact speak the truth.

## 3. AN OBJECTION AND A REPLY

But here an objection rears its ugly head. In trying to assess P(R/N&E), I suggested that semantic epiphenomenalism was very likely, given materialism, because a neural structure would have caused the same behavior if it had had *different* content but the *same* NP properties. But, says the objector, it *couldn't* have had the same NP properties but different content; having a given content *just is* having a certain set of NP properties. This is a sensible objection. Given materialism, there is a way of looking at the relation between content (as well as other mental properties) and NP properties according to which the objector is clearly right. We must therefore look a bit more deeply into that relation. Here there are fundamentally two positions: *reductionism* or *reductive materialism* on the one hand, and *non-reductive materialism* on the other. Consider the property of having as content the proposition *naturalism is all the rage these days,* and call this property "C." According to *reductive* materialism, C *just is* a certain combination of NP properties.

[20] It might be a disjunction of such properties; more likely a complex Boolean construction on NP properties, perhaps something like

$$(P_1\&P_7\&P_{28}\ldots)\ v\ (P_3\&P_{34}\&P_{17}\&\ldots)\ v\ (P_8\&P_{83}\&P_{107}\&\ldots)\ v\ \ldots$$

(where the $P_i$ are NP properties).

Now take any belief B you like: what is the probability that B is true given N&E and reductive materialism? What we know is that B has a certain content, that having that content just is having a certain combination of NP properties, and (we may assume) that having that combination of NP properties (in the circumstances in which the organism finds itself) is adaptive. What, then, is the probability that the content of B is *true?* Well, it doesn't *matter* whether it is true; if it is true, the NP properties constituting that content will be adaptive, but if it is false, those properties will be equally adaptive, since in each case they make the same causal contribution to behavior. That combination of NP properties is the property of having a certain content; it is the property of being associated with a certain proposition *p* in such a way that *p* is the content of the belief. Having that combination of NP properties is adaptive; hence having that belief is adaptive; but that combination of NP properties will be equally adaptive whether *p* is true or false. In this case (reductionism) content does enter into the causal chain leading to behavior, because NP properties do, and having a certain content just is displaying a certain set of NP properties. But those properties will be adaptive, whether or not the content the having of which they constitute is true. Content enters in, all right, but not, we might say, *as* content. Better, content enters the causal chain leading to behavior, but not in such a way that its truth or falsehood bears on the adaptive character of the belief.

But, someone might object, given that the belief is adaptive, isn't there a greater probability of its being true than of its being false? Why so? Because, the objector continues, the belief's being adaptive means that having this belief, in these or similar circumstances, helped the creature's ancestors to survive and reproduce; having this belief contributed to reproductive fitness. And wouldn't the best explanation for this contribution be that the belief accurately represented their circumstances, i.e., *was true?* So, probably, the belief was adaptive for the creature's ancestors because it was true. So, probably, the belief is adaptive for this creature in *its* circumstances because it is true.[21]

This objection, beguiling as it sounds, is mistaken. The proper explanation of this belief's being adaptive is that having the NP properties that constitute the content of the belief causes adaptive behavior, not that the belief is true. And, of course, having those NP properties can cause adaptive behavior whether or not the content they constitute is true. At a certain level of complexity of NP properties, the neural structure that displays those properties also acquires a certain content C. That is because having that particular complex of NP properties just is what it is to have C. Having those NP properties, presumably, is adaptive; but whether the content arising in this way is true or false makes no difference to that adaptivity. What explains the adaptivity is just that having these NP properties, this content, causes adaptive behavior.[22]

So consider again a belief B with its content C: what, given that having that belief is adaptive, is the probability that C is true, is a true proposition? Well, since truth of content doesn't make a difference to the adaptivity of the belief, the belief could be true, but it could equally likely be false. We'd have to estimate the probability that it is true as about 0.5. But then if the creature has 1,000 independent beliefs, the probability that, say, 3/4 of them are true (and at least this would be required for reliability) will be very low—less than $10^{-58}$.[23] So on naturalism and reductionism, the probability of R appears to be very low.

That's how things go given reductive materialism; according to *non-reductive materialism,* the other possibility, a mental property is not an NP property or any Boolean construction on NP properties, but a new sort of property that gets instantiated when a neural structure attains a certain degree of complexity, that is, when it displays a certain complex set of NP properties. (We might call it an "emergent" property.) On this way of looking at the matter, mental properties are not identical with or reducible to NP properties; they supervene on them. But again, take any particular belief B: what is the probability, on N&E & non-reductive materialism, that B is true? What we know is that B has a content, that this content arises when the structure has a certain complex set of NP properties, and that having that set of NP properties is adaptive. But once again, it doesn't matter for adaptivity whether the content associated with those NP properties is true or false; so once again, the probability that the content is true will have to be estimated as about .5; but then once again the probability that their faculties are reliable is low. Either way, therefore, the probability is low, so that P(R/N&E) is also low—or, as we could add, if we like, inscrutable. But then either way one who accepts N&E and grasps the above argument has a defeater for R.

## 4. A DEFEATER-DEFEATER?

Now of course defeaters can be themselves defeated. For example, I know that you are a lifeguard and believe on that ground that you are an excellent swimmer. But then I learn that 45 percent of Frisian lifeguards are poor swimmers, and I know that you are Frisian: this gives me a defeater for the belief that you are a fine swimmer. But then I learn still further that you graduated from the Department of Lifeguarding at the University of Leeuwarden and that one of the requirements for graduation is being an excellent swimmer: that gives me a defeater for the defeater of my original belief: a defeater-defeater as we might put it.[24] But (to return to our argument) can't the defeater the naturalist has for R be in turn defeated? Could she perhaps get a defeater-defeater by doing some science—for example, determining by scientific means that her faculties really are reliable? Couldn’t she go to the MIT cognitive-reliability laboratory for a check-up?

That won’t help; clearly that course would *presuppose* that her faculties are reliable; she’d be relying on the accuracy of her faculties in believing that there is such a thing as MIT, that she has in fact consulted its scientists, that they have given her a clean bill of cognitive health, and the like. Thomas Reid (*Essays on the Intellectual Powers of Man*) put it like this: “If a man’s honesty were called into question, it would be ridiculous to refer to the man’s own word, whether he be honest or not. The same absurdity there is in attempting to prove, by any kind of reasoning, probable or demonstrative, that our reason is not fallacious, since the very point in question is, whether reasoning may be trusted” (276).

Is there any sensible way at all in which she can argue for R? It seems not. Any argument she might produce will have premises; and these premises, she claims, give her good reason to believe R. But of course she has the very same defeater for each of those premises that she has for R, as well as for the belief that if the premises of that argument are true, then so is the conclusion. So this defeater can’t be defeated. We could also put it like this: any argument she offers for R is epistemically circular (to use William Alston’s terminology). Naturalistic evolution gives its adherents a reason for doubting that our beliefs are mostly true; perhaps they are mostly mistaken. But then it won’t help to *argue* that they can’t be mostly mistaken; for the very reason for mistrusting our cognitive faculties *generally* will be a reason for mistrusting the faculties that produce belief in the goodness of the argument.

## 5. N&E DEFEATED

He who accepts N&E, therefore, has a defeater for R—a defeater, furthermore, that can’t be defeated. But if you have a defeater for R, the belief that your cognitive faculties are reliable, then you have a defeater for any belief you take to be produced by your cognitive faculties. Naturally enough, that would be all of your beliefs. Once more, this defeater cannot be defeated; any proposed defeater would be an argument, and any such argument would be epistemically circular. Hence the devotee of N&E has a defeater for just any belief he holds—a defeater that is ultimately undefeated. But this means, obviously, that he has a defeater for N&E itself. N&E, therefore, cannot rationally be accepted—at any rate by someone who is apprised of this argument and sees the connections between N&E and R.

And that has two important consequences. The first is that there is a conflict between naturalism and science: one can’t rationally accept the conjunction of naturalism with evolution; but evolution is one of the main pillars of contemporary science. Hence there is a religion/science conflict in this area, all right, but it isn’t between Christian belief and science: it is between naturalism and science. Second, if naturalism is true, then so, in all probability, is evolution; evolution is the only game in town, for the naturalist, with respect to the question how all this variety of flora and fauna has arisen. If *that* is so, finally, then naturalism *simpliciter* is self-defeating and cannot rationally be accepted—at any rate by someone who sees the connections between N&E and R.

Notes

1. A version of "An Evolutionary Argument against Naturalism" (hereafter "EAAN") first appeared in *Logos* 12: 27–49, 1991 and in *Warrant and Proper Function* (New York: Oxford University Press, 1993; hereafter "WPF"), ch. 12. These statements of the argument contain an error: at a certain point the statement of the "preliminary argument" (WPF 228–29) confuses the unconditional objective or logical probability of R with its probability conditional on our background knowledge. (Here I was helped by Brandon Fitelson and Elliot Sober: see their paper "Plantinga's Probability Arguments against Evolutionary Naturalism" in *Pacific Philosophical Quarterly* 79 [1998], 115–29). For details and repair, see my *Warranted Christian Belief* (New York: Oxford University Press, 2000), 229ff. The main argument, happily, is unaffected. See also "Probability and Defeaters," *Pacific Philosophical Quarterly* 84:3 (Sept. 2003), my reply to Sober and Fitelson. For further critical comment on EAAN (together with my replies) see *Naturalism Defeated?* ed. James Beilby (Ithaca, NY: Cornell University Press, 2002).

2. Erroneously, in my opinion. There is no inner connection between science and naturalism; indeed, as I'm about to argue, naturalism clashes with science.

3. "*Quam Dialectica*" in *God and the Philosophers,* ed. Thomas Morris (New York: Oxford University Press, 1994), 49

4. Thus Richard Dawkins: "Although atheism might have been logically tenable before Darwin, Darwin made it possible to be an intellectually fulfilled atheist." *The Blind Watchmaker* (New York: Norton, 1986), 6–7.

5. *Very* roughly: a thermometer stuck on 72 degrees F. isn't reliable even if it is located somewhere (San Diego?) where it is 72 F. degrees nearly all of the time. What the thermometer (and our cognitive faculties) would do if things were different in certain (hard to specify) respects is also relevant. Again, if our aim were to analyze *reliability* much more would have to be said. Note that for reliability thus construed, it is not enough that the beliefs produced be fitness enhancing.

6. Thus Thomas Aquinas: "Since human beings are said to be in the image of God in virtue of their having a nature that includes an intellect, such a nature is most in the image of God in virtue of being most able to imitate God" (ST Ia q. 93 a. 4); *and* "Only in rational creatures is there found

a likeness of God which counts as an image.… As far as a likeness of the divine nature is concerned, rational creatures seem somehow to attain a representation of [that] type in virtue of imitating God not only in this, that he is and lives, but especially in this, that he understands" (ST Ia Q.93 a.6).

7. You might think not: if our origin involves *random* genetic variation, then we and our cognitive faculties would have developed by way of *chance* rather than by way of design, as would be required by our having been created by God in his image. But this is to import far too much into the biologist's term "random." Those random variations are random in the sense that they don't arise out of the organism's design plan and don't ordinarily play a role in its viability; perhaps they are also random in the sense that they are not predictable. But of course it doesn't follow that they are random in the much stronger sense of not being caused, orchestrated, or arranged by anything at all (including God). And suppose the biologists, or others, *did* intend this stronger sense of "random": then their theory (call it "T") would indeed entail that human beings have not been designed by God. But T would not be more probable than not with respect to the evidence. For there would be an empirically equivalent theory (the theory that results from T by taking the weaker sense of "random" and adding that God has orchestrated the mutations) that is inconsistent with T but is as well supported by the evidence; if so, T is not more probable than not with respect to the relevant evidence.

8. *Darwin's Cathedral: Evolution, Religion and the Nature of Society* (Chicago: University of Chicago Press, 2002), 228.

9. Letter to William Graham, Down, July 3, 1881. In *The Life and Letters of Charles Darwin Including an Autobiographical Chapter,* ed. Francis Darwin (London: John Murray, Albermarle Street, 1887), vol. 1, 315–16. Evan Fales and Omar Mirza point out that Darwin probably had in mind, here, not everyday beliefs such as that the teapot is in the cupboard, but religious and philosophical convictions—naturalism, for example.

10. *Journal of Philosophy* 84 (Oct. 1987), 548.

11. For an account of objective probability, see my *Warrant and Proper Function* (New York: Oxford University Press, 1993), 161ff.

12. Alternatively, we might say that the probability here is inscrutable, such that we can't make an estimate of it. It is unlikely that a large set of beliefs (comparable in size to the number of beliefs a human being has) should contain mainly truths; that gives us a reason for regarding the probability in question as low. On the other hand, we know something further about the relevant set of propositions, namely, that it is a set each member of which is believed by someone. How

does this affect the probability? Here perhaps we don't know what to say, and should conclude that the probability in question is inscrutable. (Here I am indebted to John Hare.)

13. Must we concur with Donald Davidson, who thinks it is "impossible correctly to hold that anyone could be mostly wrong about how things are"? ("A Coherence Theory of Truth and Knowledge" in *Kant oder Hegel?* ed. Dieter Henrich [Stuttgart: Klett-Cotta Buchhandlung, 1983], 535.) No; what Davidson shows (if anything) is that it isn't possible for me to *understand* another creature, unless I suppose that she holds mainly true beliefs. That may (or more likely, may not) be so; but it doesn't follow that there couldn't be creatures with mainly false beliefs, and *a fortiori* it doesn't follow that my own beliefs are mainly true. Davidson went on to argue that an *omniscient* interpreter would have to use the same methods we have to use and would therefore have to suppose her interlocutor held mostly true beliefs; given the omniscient interpreter's omniscience, he concluded that her interlocutor would in fact have mostly true beliefs. In so concluding, however, he apparently employs the premise that any proposition that would be believed by any omniscient being is true; this premise directly yields the conclusion that there *is* an omniscient being (since any omniscient being worth its salt will believe that there is an omniscient being), a conclusion to which Davidson might not have wished to commit himself. See WPF 80–81.

14. First so-called by T. H. Huxley, ("Darwin's bulldog"): "It may be be assumed … that molecular changes in the brain are the causes of all the states of consciousness.… [But is] there any evidence that these stages of consciousness may, conversely, cause … molecular changes [in the brain] which give rise to muscular motion? I see no such evidence … [Consciousness appears] to be … completely without any power of modifying [the] working of the body, just as the steam whistle … of a locomotive engine is without influence upon its machinery." ("On the Hypothesis that Animals are Automata and its History" [1874], chapter 5 of his *Method and Results* [London: Macmillan, 1893], 239–40.) Later in the essay: "To the best of my judgment, the argumentation which applies to brutes holds equally good of men; and therefore … all states of consciousness in us, as in them, are immediately caused by molecular changes of the brain-substance. It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of the organism.… We are conscious automata . . ." (243–44). (Note the occurrence here of that widely endorsed form of argument, "I know of no proof that not-*p;* therefore there is no proof that not-*p;* therefore *p.*") However, I am here using the term to denote *any* view according to which belief isn't involved in the causal chain leading to behavior, whether or not that view involves the dualism that seems to be part of Huxley's version.

15. *Meaning and Mental Representation* (Cambridge, MA: MIT Press, 1989), 130. In *Explaining Behavior* (Cambridge, MA: MIT Press, 1988) Fred Dretske makes a valiant (but in my opinion unsuccessful) effort to explain how, given materialism about human beings, it could be

that beliefs (and other representations) play a causal role in the production of behavior by virtue of their content or semantics.

16. *Meaning and Mental Representation* (Cambridge, MA: MIT Press, 1989), 130.

17. Fred Dretske's *Explaining Behavior* (Cambridge, MA: MIT Press, 1988) is surely one of the most promising of these attempts; it fails, however (as I see it), among other things, because it implies that there are no distinct but logically equivalent beliefs, and indeed no distinct but *causally* equivalent beliefs.

18. Of course these figures are the merest approximations; others might make the estimates somewhat differently; but they can be significantly altered without significantly altering the final result. For example, perhaps you think the P(R/N&C) is higher, perhaps even 1; then (retaining the other assignments) P(R/N) will be in the neighborhood of 0.44. Or perhaps you reject the thought that P(-C/N) is more probable than P(C/N), thinking them about equal. Then (again, retaining the other assignments) P(R/N) will be in the neighborhood of 0.55.

19. I shall use this term to mean "failing to believe," so that I withhold p if either I believe its denial or I believe neither it nor its denial.

20. Or (to accommodate the thought that meaning "ain't in the head") a combination of NP properties with environmental properties. I'll assume but not mention this qualification in what follows.

21. Here I am indebted to Tom Crisp.

22. In this connection, consider dream beliefs. Take a given dream belief with its content C: Having the NP properties that constitute the property of having C is adaptive; but it makes no difference whether or not that content is true.

23. As calculated by the eminent mathematician Paul Zwier. This is the probability that the whole battery of cognitive faculties is reliable; the probability that a given faculty is reliable will be larger, but still small; if its output is, say, 100 beliefs, the probability that 3/4 of them are true will be no more than 0.000001.

24. As in fact John Pollock *does* put it.