AN EVOLUTIONARY ARGUMENT AGAINST NATURALISM

Alvin Plantinga

UDK 141.111

The Problem

Most of us think (or would think on reflection) that at least *a* function or purpose of our cognitive faculties is to provide us with true beliefs. Moreover, we go on to think that when they function properly, in accord with our design plan, then for the most part they do precisely that. Of course qualifications are necessary. There are various exceptions and special cases: visual illusions, mechanisms like forgetting the pain of childbirth, optimism about recovery not warranted by the relevant statistics, unintended conceptual byproducts, and so on. There are also those areas of cognitive endeavor marked by enormous disagreement, wildly varying opinion: philosophy and Scripture scholarship come to mind. Here the sheer volume of disagreement and the great variety and contrariety of options proposed suggest that either not all of us are such that our cognitive faculties *do* function according to the design plan, in these areas, or that it is not the case that the relevant modules of the design plan are aimed at truth, or that the design plan for those areas is defective.

Nevertheless over a vast area of cognitive terrain we take it both that the purpose (function) of our cognitive faculties is to provide us with true or verisimilitudinous beliefs, and that, for the most part, that is just what they do. We think our faculties much better adapted to reach the truth in some areas than others; we are good at elementary arithmetic and logic, and the perception of middle–sized objects under ordinary conditions. We are also good at remembering certain sorts of things: I can easily remember what I had for breakfast this morning, where my office was located yesterday, and whether there was a large explosion in my house last night. Things get more difficult, however, when it comes to an accurate reconstruction of what it was like to be, say, a fifth century BC Greek (not to mention a bat), or whether the axiom of choice or the continuum hypothesis is true; things are even more difficult, perhaps, when it comes to figuring out how quantum mechanics is to be understood, and what the subnuclear realm of quark and gluon is really like, if

indeed there really is a subnuclear realm of quark and gluon. Still, there remains a vast portion of our cognitive terrain where we think that our cognitive faculties do furnish us with truth.

But isn't there a problem, here, for the naturalist? It's a little difficult to say precisely what naturalism is, but at any rate it entails there is no such person as God, or anything at all like God. Many naturalists go on to add that the only things that exist are the things postulated or admitted by natural science. At any rate for the naturalist who thinks that we and our cognitive capacities arrived upon the scene after some billions of years of evolution (by way of natural selection, genetic drift, and other blind processes working on such sources of genetic variation as random genetic mutation)? Richard Dawkins (according to Peter Medawar, "one of the most brilliant of the rising generation of biologists") once leaned over and remarked to A. J. Ayer at one of those elegant, candle–lit, bibulous Oxford college dinners that he couldn't imagine being an atheist before 1859 (the year Darwin's *Origin of Species* was published); "... although atheism might have been logically tenable before Darwin", said he, "Darwin made it possible to be an intellectually fulfilled atheist".[1]

Now Dawkins thinks Darwin made it possible to be an intellectually fulfilled atheist. But perhaps Dawkins is dead wrong here. Perhaps the truth lies in the opposite direction. If our cognitive faculties have originated as Dawkins thinks, then their ultimate purpose or function (if they *have* a purpose or function) will be something like *survival* (of individual, species, gene or genotype); but then it seems initially doubtful that among their functions — ultimate, proximate, or otherwise — would be the production of true beliefs. Taking up this theme, Patricia Churchland declares that the most important thing about the human brain is that it has evolved; hence, she says, 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.[2]

1 Dawkins, R. (1986), The Blind Watchmaker. London and New York: W. W. Norton & Co., 6–7.

2 Churchland, P. S. (1987), Epistemology in the Age of Neuroscience, in: Journal of Philosophy 84, 548.

Her point, I think, is that (from a naturalistic perspective) what evolution guarantees is (at most) that we *behave* in certain ways — in such ways as to promote survival, or survival through childbearing age, or, more exactly, fitness. The principal function or purpose, then, (the 'chore' says Churchland) of our cognitive faculties is not that of producing true or verisimilitudinous beliefs, but instead that of contributing to survival by getting the body parts in the right place. What evolution underwrites is only (at most) that our *behavior* be reasonably adaptive to the circumstances in which our ancestors found themselves; hence (so far forth) it does not guarantee mostly true or verisimilitudinous beliefs. Of course our beliefs *might* be mostly true or verisimilitudinous (hereafter I'll omit the 'versimilitudinous'); but there is no particular reason to think they *would* be: natural selection is interested, not in truth, but in appropriate behavior. What Churchland says suggests, therefore, that naturalistic evolution — that is, the conjunction of metaphysical naturalism with the view that we and our cognitive faculties have arisen by way of the mechanisms and processes proposed by contemporary evolutionary theory — gives us reason to doubt two things: (a) that a *purpose* of our cognitive systems is that of serving us with true beliefs, and (b) that they *do*, in fact, furnish us with mostly true beliefs.

W. V. O. Quine and Karl Popper, however, apparently demur. Popper argues that since we have evolved and survived, we may be pretty sure that our hypotheses and guesses as to what the world is like are mostly correct.[3] And Quine says he finds encouragement in Darwin:

> What does make clear sense is this other part of the problem of induction: why does our innate subjective spacing of qualities accord so well with the functionally relevant groupings in nature as to make our inductions tend to come out right? Why should our subjective spacing of qualities have a special purchase on nature and a lien on the future?
>
> There is some encouragement in Darwin. If people's innate spacing of qualities is a gene–linked trait, then the spacing that has made for the most successful inductions will have tended to predominate through natural selection. Creatures inveterately wrong in their inductions have a pathetic but praiseworthy tendency to die before reproducing their kind.[4]

Indeed, Quine finds a great deal more encouragement in Darwin than Darwin did: "With me," says Darwin,

3 Popper, K. R. (1972), Objective Knowledge. An Evolutionary Approach. Oxford: Clarendon Press, 261.

4 Quine, W. V. O. (1969), Natural Kinds, in: Ontological Relativity and Other Essays. New York: Columbia University Press, 126.

> 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?[5]

So here we appear to have Quine and Popper on one side and Darwin and Churchland on the other. Who is right? But a prior question: what, precisely, is the issue? Darwin and Churchland seem to believe that (naturalistic) evolution gives one a reason to doubt that human cognitive faculties produce for the most part true beliefs: call this 'Darwin's Doubt'. Quine and Popper, on the other hand, apparently hold that evolution gives us reason to believe the opposite: that human cognitive faculties *do* produce for the most part true beliefs. How shall we understand this opposition?

Darwin's Doubt

One possibility: perhaps Darwin and Churchland mean to propose that a certain objective conditional probability is relatively low: the probability of human cognitive faculties' being reliable (producing mostly true beliefs), given that human beings *have* cognitive faculties (of the sort we have) and given that these faculties have been produced by evolution (Dawkin's blind evolution, unguided by the hand of God or any other person). If metaphysical naturalism and this evolutionary account are both true, then our cognitive faculties will have resulted from blind mechanisms like natural selection, working on such sources of genetic variation as random genetic mutation. Evolution is interested, not in true belief, but survival (or fitness). It is therefore unlikely that our cognitive faculties have the production of true belief as a proximate or any other function, and the probability of our faculties' being reliable, (given naturalistic evolution) would be fairly low. Popper and Quine, on the other side, judge that probability fairly high.

The issue, then, is the value of a certain conditional probability: P(R/N&E).[6] Here N is metaphysical naturalism. It isn't easy to say precisely what naturalism *is*, but perhaps that isn't necessary in this context; prominent

5 Darwin, Ch. (1887), Letter to William Graham, Down, July 3rd, 1881, in: The Life and Letters of Charles Darwin, Including an Autobiographical Chapter, Volume 1, ed. Francis Darwin. London: John Murray, Albermarle Street, 315–316.

6 We could think of this probability in two ways: as a conditional epistemic probability, or as a conditional objective probability. Either will serve for my argument, but I should think the better way to think of it would be as objective probability; for in this sort of context epistemic probability, presumably, should follow known (or conjectured) objective probability.

examples would be the views of (say) David Armstrong, the later Darwin, Quine, and Bertrand Russell. Crucial to metaphysical naturalism, of course, is the view that there is no such person as the God of traditional theism; and many naturalists hold that what exists are the entities postulated in science together with things composed of those entities. E is the proposition that human cognitive faculties arose by way of the mechanisms to which contemporary evolutionary thought directs our attention. R, on the other hand, is the claim that our cognitive faculties are reliable (on the whole, and with the qualifications mentioned above), in the sense that they produce mostly true beliefs in the sorts of environments that are normal for them. And the question is: what is the probability of R on N&E? (Alternatively, perhaps the interest of *that* question lies in its bearing on *this* question: what is the probability that a belief produced by human cognitive faculties is *true*, given N&E?) And if we construe the dispute in this way, then what Darwin and Churchland propose is that this probability is relatively low, while Quine and Popper think it fairly high.

Let's see if we can develop Darwin's doubt. In order to avoid interspecific chauvinism, suppose we think, first, not about ourselves and our ancestors, but about a hypothetical population of creatures a lot like ourselves on a planet similar to Earth. (Darwin proposed that we think about another species, such as monkeys.) Suppose these creatures have cognitive faculties, hold beliefs, change beliefs, make inferences, and so on; and suppose they have arisen by way of the selection processes endorsed by contemporary evolutionary thought. What is the probability that their faculties are reliable? What is P(R/N&E), specified not to us, but to them? According to Quine and Popper, the probability in question would be rather high: belief is connected with action in such a way that extensive false belief would lead to maladaptive behavior, in which case it is likely that the ancestors of those creatures would have displayed that pathetic but praiseworthy tendency Quine mentions.

But now for the contrary argument. First, perhaps it is likely that their *behavior* is adaptive; but nothing so far follows about their *beliefs*. And in fact natural selection will be able to mold their beliefs in the direction of truth only if there is the appropriate relation between belief and behavior. What are the possibilities here?

(1) One possibility is *epiphenomenalism*:[7] their behavior is not caused by their beliefs. On this possibility, their movement and behavior would be caused by

7 First so–called by T. H. Huxley, ("Darwin's bull dog"): "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 states of consciousness may, conversely, cause [...] molecular changes [in the brain] which give rise to muscular motion?" I see no such evidence [...]

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, it will be rather unlikely that their beliefs display the great preponderance of truth over falsehood required by their cognitive faculties' being reliable; that is because a randomly chosen large set of propositions is unlikely to display such a preponderance of truth over falsehood. 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*. Put in currently fashionable jargon, this would be the suggestion that beliefs are indeed causally efficacious, but by virtue of their *syntax*, not by virtue of their *semantics*. On a naturalist or anyway a materialist way of thinking, a belief would perhaps be something like a long–term pattern of neural activity, a long–term neuronal event. This event will have properties of at least two different kinds. On the one hand, there are its neuro-

[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."(Huxley, T. H. (1874), On the Hypothesis that Animals are Automate and its History (= chapter 5 of his Method and Results. London: Macmillan 1893, 239–240)). 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 [...]" (243–244; 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').

physiological or electrochemical properties: the number of neurons involved in the belief, the connections between them, their firing thresholds, the rate and strength at which they fire, the way in which these change over time and in response to other neural activity, and so on. Call these *syntactical* properties of the belief. On the other hand, however, if the belief is really a *belief*, it will be the belief that *p* for some proposition *p*. Perhaps it is the belief that there once was a brewery where the Metropolitan Opera House now stands. This proposition, we might say, is the *content* of the belief in question. So in addition to its syntactical properties, a belief will also have *semantical*[8] properties — for example, the property of being the belief that there once was a brewery where the Metropolitan Opera House now stands. (Other semantical properties: being true or false, *entailing that there has been at least one brewery*, *being consistent with the proposition that all men are mortal* and so on.) And the second possibility is that 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 (if grudgingly) popular among contemporary philosophers of mind; indeed, Robert Cummins goes so far as to call it the "received view."[9]

The probability of R on N&E together with this possibility will also be relatively low. The reason is that truth or 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.[10] But then it will be unlikely that their beliefs

8 Granted: the analogies between these properties and syntax and semantics is a bit distant and strained; here I am just following current custom.

9 Cummins, R. (1989), Meaning and Mental Representation. Cambridge / Mass.: MIT Press, 130. — In Explaining Behavior (Cambridge / Mass.: 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.

10 We must also consider here the possibility that the syntax and semantics of belief are the effects of a common cause: perhaps there is a cause of a belief's having certain adaptive syntactic properties, which also causes the belief to have the semantic properties it does (it brings it about that the event in question is the belief that p for some proposition p); and perhaps this cause brings it about that a true proposition is associated with the belief (the neuronal event) in question. (Here I was instructed by William Ramsey and Patrick Kain.) What would be the likelihood, given N&E, that there is such a common cause at work? I suppose it would be relatively low: why should this common cause associate true propositions with these neuronal events? But perhaps the right answer is not that the probability in question is low, but that it is inscrutable: see below, pp. 49ff.

display the preponderance of truth over falsehood required by their cognitive faculties being reliable.

(3) It could be that beliefs are causally efficacious — 'semantically' as well as 'syntactically' — with respect to behavior, 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 causally connected with their behavior — by way of content as well as neurophysiological properties — and also adaptive. (I suppose this is the common sense view of the connection between behavior and belief in our own case.) What is the probability (on this assumption 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? 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; but this behavior could be produced by a large number of different belief–desire pairs. 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. 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 this last argument 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* or *all* of his beliefs might 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. In these cases, his beliefs would all be false; but this is 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.

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 (both syntactically and semantically) with respect to their behavior but maladaptive, and the possibility that their beliefs are causally efficacious (both syntactically and semantically) with respect to behavior and adaptive. P(R/N&E) will be the weighted average of P(R/N& E&Pi) 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:

$$\begin{aligned} P(R/N\&E) = {} & (P(R/N\&E\&P_1) \times P(P_1/N\&E)) \\ & + (P(R/N\&E\&P_2) \times P(P_2/N\&E)) \\ & + (P(R/N\&E\&P_3) \times P(P_3/N\&E)) \\ & + (P(R/N\&E\&P_4) \times P(P_4/N\&E)). \end{aligned}$$

But we can simplify. Note that epiphenomenalism *simpliciter* 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 I have argued so far seen is that the probability of R on N&–C is low; that seems reasonable. But perhaps we could instead say that this probability is *inscrutable* i. e., such that we simply cannot determine what this probability is. Similarly, what we should say about the probability of R on N&C is that it is either moderately high or else inscrutable.

And of course what we are looking for is P(R/N&E). Since C and –C are jointly exhaustive and mutually exclusive, we can write

$$P(R/N\&E) = P(R/N\&E\&C) \times P(C/N\&E) + P(R/N\&E\&–C) \times P(–C/N\&E)$$

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.

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 ordinary naturalism: what is the probability that one or the other of the two epiphenomenalistic scenarios is true? Note that according to Robert Cummins, semantic epiphenomenalism is in fact the received view as to the relation between belief and behavior.[11] 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. If a belief just is a neural structure of some kind — a structure that somehow possesses content — then it is exceedingly hard to see how its content (as opposed to its neurophysiological properties) can get involved in the causal chain leading to behavior: had a given such structure had a different content but the same neurophysiological properties, its causal contribution to behavior, one thinks, would be the same. On the other hand, if a belief is not a material structure at all, but a nonphysical bit of consciousness, it is hard to see that there is any room for it in the causal chain leading to behavior; what causes the muscular contractions involved in behavior will be states of the nervous system, with no point at which this nonphysical bit of consciousness makes a causal contribution. So it is exceedingly hard to see, given N&E, how the content of a belief can have causal efficacy.

It is exceedingly hard to see, that is, how epiphenomenalism — semantic or *simpliciter* — can be avoided, given N&E. (There have been some valiant efforts[12] 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 of course P(C/N&E) will be 0. 3. But of course we could easily be wrong;

11 Cummins, R. (1989), Meaning and Mental Representation. Cambridge / Mass.: MIT Press, 130

12 Fred Dretske's Explaining Behavior (Cambridge / Mass.: 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.

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 really 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. 9. And given these estimates, P(R/N) will be in the neighborhood of 0.41.[13] 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 relatively low — less than 0. 5, at any rate — or inscrutable.

The Argument

What I have argued so far is that P(R/N&E) is either low or inscrutable. What I shall argue next is that in either case one who accepts N&E (and sees that this probability is low or inscrutable) has a *defeater* for R: a reason to reject it, to withhold it. I will argue this by turning to three analogies.

Compare first the case of a believer in God, who, perhaps through an injudicious reading of Freud, comes to think that religious belief generally and theistic belief in particular is almost always produced by wish fulfillment. Such beliefs, she now thinks, are not produced by cognitive faculties functioning properly in a congenial environment according to a design plan successfully aimed at truth; instead they are produced by wish fulfillment, which, while indeed it has a function, does not have the function of producing true beliefs. Suppose she considers the objective probability that wish fulfillment, as a belief–producing mechanism, is reliable. She might quite properly

13 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.

estimate this probability as relatively low; alternatively, however, she might think the right course, here, is agnosticism; and she might also be equally agnostic about the probability that a belief should be true, given that it is produced by wish fulfillment.

But then in either case she has a defeater for any belief she takes to be produced by the mechanism in question. Consider the first case: she thinks the probability that wish fulfillment is reliable is low, and the probability that a belief should be true, given that it is produced by wish fulfillment not far from. 5. Then she clearly has a straightforward defeater for any of her beliefs she takes to be produced by wish fulfillment, including belief in God. But the same holds if she takes this probability — i. e., the probability that a belief is true, given that it is produced by wish fulfillment — to be inscrutable: in that case too she clearly has a defeater for any belief she takes to be produced by that belief–producing process. (Her position is like that of someone who buys a thermometer, believing it to be reliable, but then learns that this instrument was produced in a factory owned by an eccentric who mixes in a number of unreliable thermometers into the factory's output. If she doesn't have any idea as to the proportion of unreliable thermometers in the factory's output, she will have a defeater for her initial belief that the thermometer is reliable.) Either way, therefore, she has a defeater for her belief in God.

Second analogy: suppose a person comes into a factory and sees an assembly line carrying apparently red widgets; suppose the widgets look red. Naturally enough, she forms the belief that x, the widget she is looking at, is red. The shop superintendent, however, tells her that these widgets are being irradiated by a variety of red light that makes it possible to detect otherwise undetectable hairline cracks. Then she has a defeater for her initial belief that x is red. To use John Pollock's terminology (and since I am already filching his example, why not?) she has an *undercutting* defeater (rather than a *rebutting* defeater). It isn't that she has acquired some evidence for that widget's being nonred, thus rebutting the belief that it is red; it is rather that her grounds for thinking it red have been undercut. And, indeed, upon hearing (and believing) that the widgets are being thus irradiated, she will no longer believe that the widget in question is red.

Consider, on the other hand, a second kind of case. As before, the shop superintendent tells her that those widgets are being irradiated by red light; but then a vice president comes along and tells her that the shop superintendent suffers from a highly resilient but fortunately specific hallucination, so that he is reliable on other topics even if totally unreliable on red lights and widgets. Still, the vice president *himself* doesn't look wholly reliable: there is a certain shiftiness about the eyes... Then she doesn't know *what* to believe about those alleged red lights. What will she properly think about the color of the widgets? Here she doesn't come to believe that the probability of a wid-

get's being red, given that it looks red, is low; instead, she is agnostic about that probability; she won't know what that probability might be; for all she knows it could be very low, but also, for all she knows, it could be high. The rational course for her, therefore, is to be agnostic about the deliverances of her visual perception (so far as color detection is concerned) in this situation. But then she also has a defeater for initial belief that x is red.

A third analogy: suppose I come to 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 brain–in–vat version of this scenario, and suppose I come to believe that I have been captured by Alpha–Centaurian superscientists 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, I 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[14] my natural belief that my cognitive faculties are in fact reliable.

But now suppose we return to the person convinced of N&E (and who sees that P(R/N&E) is low or inscrutable). He is in the same position with respect to R as is the above believer in God. He is in the same condition, with respect to R, as the widget observer, and as the person who thinks he is a victim of a Cartesian evil demon or an Alpha–Centaurian superscientist. So he too has a defeater for R, a good reason for being agnostic with respect to it. If he has no defeater for that defeater, and no other source of evidence, the right attitude towards R would be agnosticism. That is not to say that he would in fact be able to reject R. Due to that animal faith noted by Hume, Reid, and Santayana (but so–called only by the latter), chances are he wouldn't; still, agnosticism is what reason requires. One who accepts N&E and sees that P(R/N&E) is low or inscrutable has a defeater for R.

By way of brief review: Darwin's doubt can be taken as the claim that the probability of R on N&E is fairly low; as I argued above, that is plausible. But Darwin's doubt can also be taken as the claim that the rational attitude to take,

14 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.

here, is agnosticism about that probability; that is more plausible. Still more plausible is the disjunction of these two claims: either the rational attitude to take towards this probability is the judgment that it is low, or the rational attitude is agnosticism with respect to it. But then the devotee of N&E has a defeater for R.

If I have an undefeated defeater for R, however, then by the same token I have an undefeated defeater for any other belief *B* my cognitive faculties produce, a reason to be doubtful of that belief, a reason to withhold it. For any such belief will be produced by cognitive faculties that I cannot rationally believe to be reliable. But then clearly the same will be true for any proposition they produce: the fact that I can't rationally believe that the faculties that produce that belief are reliable, gives me a reason for rejecting the belief. So the devotee of N&E has a defeater for just any belief *B* he holds. Now the next thing to note is that *B might be N&E itself*; our devotee of N&E has an undercutting defeater for N&E, a reason to doubt it, a reason to be agnostic with respect to it. If he has no defeater for this defeater and no independent evidence — if his reason for doubting N&E remains undefeated — then the rational course would be to reject belief in N&E. N&E is self–defeating.

And here we must note something special about N&E. So far, we have been lumping together all of our cognitive faculties, all of our sources of belief, and all the sorts of beliefs they produce. But perhaps these different sorts of faculties should be treated differently; clearly the argument can be narrowed down to specific faculties or powers or belief–producing mechanisms, with possibly different results for different cases. And surely the argument does apply more plausibly to some cognitive powers than to others. If there are such differences among those faculties or powers, presumably *perception* and *memory* would be at an advantage as compared to the cognitive mechanisms whereby we come to such beliefs as, say, that arithmetic is incomplete and the continuum hypothesis is independent of ordinary set theory. For even if we evaluated the probabilities differently from the way I suggested above, even if we thought it likely, on balance, that evolution would select for reliable cognitive faculties, this would be so only for cognitive mechanisms producing beliefs relevant to survival and reproduction. It would not hold, for example, for the mechanisms producing the beliefs involved in a logic or mathematics or set theory course. According to Fodor, "Darwinian selection guarantees that organisms either know the elements of logic or become posthumous;" but this would hold at most[15] for the most elementary bits of logic. (It is only the

15 "At most" because, as I argued above, if Darwinian selection guarantees anything, it is only that the organism's behavior is adaptive: there isn't anything in particular it needs to believe (or, a fortiori, to know).

occasional assistant professor of logic who needs to know even that first–order logic is complete in order to survive and reproduce.)

Indeed, the same would go generally for the more theoretical parts of science.[16]

> "[...] Evolution suggests a status for the distinctions we naturally make, that removes them far from the role of fundamental categories in scientific description. Classification by colour, or currently stable animal–mating groups is crucial to our survival amidst the dangers of poison and fang. This story suggests that the ability to track directly certain classes and divisions in the world is not a factor that guides scientists in theory choice. For there is no such close connection between the jungle and the blackboard. The evolutionary story clearly entails that such abilities of discrimination were 'selected for', by a filtering process that has nothing to do with successful theory choice in general. Indeed, no faculty of spontaneous discrimination can plausibly be attributed a different status within the scientific account of our evolution. Even if successful theory choice will in the future aid survival of the human race, it cannot be a trait 'selected for' already in our biological history."[17]

So even if you think Darwinian selection would make it probable that certain belief–producing mechanisms — those involved in the production of beliefs relevant to survival — are reliable, that would not hold for the mechanisms involved in the production of the theoretical claims of science — such beliefs, for example as E, the evolutionary story itself. And of course the same would go for N.

What we have seen so far, therefore, is that the devotee of N&E has a defeater for any belief he holds, and a stronger defeater for N&E itself. If he has no defeater for this defeater, and no independent evidence, then the rational attitude towards N&E would be one of agnosticism.

16 This hasn't been lost on those who have thought about the matter. According to Erwin Schrödinger, the fact that we human beings can discover the laws of nature is a marvel "that may well be beyond human understanding" (What is Life? Cambridge: University of Cambridge Press 1945, 31). According to Eugene Wigner, "The enormous usefulness of mathematics in the natural sciences is something bordering on the mysterious, and there is no rational explanation for it" (The Unreasonable Effectiveness of Mathematics in the Natural Sciences, in: Communications on Pure and Applied Mathematics 13 (1960), 2) and "It is difficult to avoid the impression that a miracle confronts us here, quite comparable in its striking nature to the miracle that the human mind can string a thousand arguments together without getting itself into contradictions, or to the two miracles of the existence of laws of nature and of the human mind's capacity to divine them" (loc. cit., 7). And Albert Einstein thought the intelligibility of the world a "miracle or an eternal mystery" (Lettres à Maurice Solovine. Paris: Gauthier — Villars 1956, 115).

17 van Fraassen, B. C. (l989), Laws and Symmetry. Oxford: Clarendon Press, 52–53.

But perhaps he will claim to have independent evidence. "True," he says, "if N&E were all I had to go on, then the right cognitive stance would be agnosticism about R and in fact about any proposition produced by my belief–producing faculties, including N&E itself. But why can't I reason inductively as follows? My cognitive faculties must indeed be reliable. For consider A1, any of my beliefs. Naturally enough, I believe A1; that is, I believe that A1 is true. So A1 is one of my beliefs and A1 is true; A2 is one of my beliefs and A2 is true, A3 is one of my beliefs and A3 is true, and so on. So, by induction, I argue that all or nearly all of my beliefs are true; I therefore conclude that my faculties are probably reliable (or at any rate probably reliable now) because as a matter of fact it is probable that each of the beliefs they have presently produced is true."

This argument ought to meet with less than universal acclaim. The friend of N&E does no better, arguing this way, than the theist who argues that wish fulfillment must be a reliable belief–producing mechanism by running a similar argument with respect to the beliefs he holds that he thinks are produced by wish fulfillment. He does no better than the widget observer who by virtue of a similar argument continues to believe that those widgets are red, even after having been told by the building superintendent that they are irradiated by red light. Clearly this is not the method of true philosophy.

Accordingly, the friend of N&E can't argue in this way that he has independent evidence for R. Of course she isn't likely to argue in *that* way; she is more likely to suggest that we consult the scientific results on the matter: what does science tell us about the likelihood that our cognitive faculties are reliable? But this can't work either. For consider any argument from science (or anywhere else) he might produce. This argument will have premises; and these premises, he claims, give him good reason to believe R (or N&E). But note that he has the very same defeater for each of those premises that he has for R and for N&E; and he has the same defeater for his belief that those premises constitute a good reason for R (or N&E). For that belief, and for each of the premises, he has a reason for doubting it, a reason for being agnostic with respect to it. This reason, obviously, cannot be defeated by an ultimately undefeated defeater. For every defeater of this reason he might have, he knows that he has a defeater–defeater: the very undercutting defeater that attached itself to R and to N&E in the first place.

We could also put it like this: any argument he offers, for R, is in this context delicately circular or question begging. It is not *formally* circular; its conclusion does not appear among its premises. It is instead (we might say) *pragmatically* circular in that it purports to give a reason for trusting our cognitive faculties, but is itself trustworthy only if those faculties (at least the ones involved in its production) are indeed trustworthy. In following this procedure and giving this argument, therefore, he subtly assumes the very propo-

sition he proposes to argue for. Once I come to doubt the reliability of my cognitive faculties, I can't properly try to allay that doubt by producing an *argument*; for in so doing I rely on the very faculties I am doubting. The conjunction of evolution with naturalism gives its adherents a reason for doubting that our beliefs are mostly true; perhaps they are mostly wildly mistaken. But then it won't help to *argue* that they can't be wildly mistaken; for the very reason for mistrusting our cognitive faculties generally will be a reason for mistrusting the faculties generating the beliefs involved in the argument.

But (someone might say) isn't there a problem with this argument for pragmatic circularity? The devotee of N&E begins, (naturally enough) by accepting N&E; upon being apprised of the above argument (so I say), he comes to see that he has a defeater for R and hence a reason for doubting N&E; hence (so I say) it is irrational for him to accept N&E, unless he has other evidence; but any purported other evidence will be subject to the same defeater as N&E. But now comes the rejoinder: as soon as our devotee of N&E comes to doubt R, he should also come to doubt his *defeater* for R; for that defeater, after all, depends upon his beliefs, which are a product of his cognitive faculties. So his defeater for R (and N&E) is also a defeater for that defeater, i. e., for *itself*. But then when he notes *that*, and *doubts* his defeater for R, he no longer *has* a defeater (undefeated or otherwise) for N&E; so how is it irrational for him to accept N&E?

What we seem to have here is one of those nasty dialectical loops to which Hume calls our attention:

> "[...] The sceptical reasonings, were it possible for them to exist, and were they not destroy'd by their subtlety, wou'd be successively both strong and weak, according to the successive dispositions of the mind. Reason first appears in possession of the throne, prescribing laws, and imposing maxims, with an absolute sway and authority. Her enemy, therefore, is oblig'd to take shelter under her protection, and by making use of rational arguments to prove the fallaciousness and imbecility of reason, produces in a manner, a patent under her hand and seal. This patent has at first an authority, proportion'd to the present and immediate authority of reason, from which it is deriv'd. But as it is suppos'd to be contradictory to reason, it gradually diminishes the force of that governing power, and its own at the same time; till at last they both vanish away into nothing, by a regular and just diminution. [...] 'Tis happy, therefore, that nature breaks the force of all sceptical arguments in time, and keeps them from having any considerable influence on the understanding."[18]

18 Hume, D., A Treatise of Human Nature. Edited with an analytical index by L. A. Selby–Bigge. Oxford: Clarendon Press, first edition 1888) Book I, Part IV, Section I, 187.

Following Hume, we might think that when the devotee of N&E notes that he has a defeater for R, then at that stage he also notes (if apprised of the present argument) that he has a defeater for N&E; indeed, he notes that he has a defeater for anything he believes. Since, however, his having a defeater for N&E depends upon some of his beliefs, what he now notes is that he has a defeater for his defeater of R and N&E; so now he no longer *has* that defeater for R and N&E. So then his original condition of believing R and assuming N&E reasserts itself: at which point he again has a defeater for R and N&E. But then he notes that *that* defeater is also a defeater of the defeater of R and N&E; hence .... So goes the paralyzing dialectic. After a few trips around this loop, we may be excused for throwing up our hands in despair, or disgust, and joining Hume in a game of backgammon.

Alternatively: perhaps the way to think about the matter is as follows. N&E (together with the proposition that P(R/N&E) is low or inscrutable) provides a defeater for R and hence for everything else believed by the devotee of N&E. The devotee of N&E therefore has a defeater for everything he believes. If so, he also has a defeater for that defeater, since it is one of his beliefs. That is indeed true; but it doesn't mean that he no longer has a defeater for R. For as long as he believes N&E (and that P(R/N&E) is low or inscrutable) he has a defeater for everything he believes — including, of course, R. The point remains, therefore: one who accepts N&E (and is apprised of the present argument) has a defeater for N&E, a defeater that cannot be ultimately defeated. And isn't it irrational to accept a belief for which you know you have a defeater that can't be ultimately defeated?

If you accept N&E, therefore, you have an ultimately undefeated reason for rejecting N&E: but then the rational thing to do is to reject N&E. If, furthermore, one also accepts the conditional *if N is true, then so is E*, one has an ultimately undefeated defeater for N. One who contemplates accepting N, and is torn, let's say, between N and theism, should reason as follows: if I were to accept N, I would have good and ultimately undefeated reason to be agnostic about N; so I shouldn't accept it. For all this argument shows, naturalism might still be true. The argument is not an argument for the falsehood of naturalism but instead for the conclusion that (for one who is aware of the present argument) accepting naturalism is irrational. It is like the self–referential argument against classical foundationalism: classical foundationalism is either false or such that I would be unjustified in accepting it; so (given that I am aware of this fact) I can't justifiably accept it.[19] But of course it doesn't follow that classical foundationalism isn't *true*; for all this argument shows, it could

19 See my Rationality and Belief in God, in: Faith and Rationality. Ed. A Plantinga and N. Wolterstorff. Notre Dame: University of Notre Dame Press 1983, 60–63.

be true, though not rationally acceptable. Similarly here; the argument isn't for the falsehood of naturalism, but for the irrationality of accepting it. The conclusion to be drawn, therefore, is that the conjunction of naturalism with evolutionary theory is self–defeating: it provides for itself an undefeated defeater. Evolution, therefore, presents naturalism with an undefeated defeater. But if naturalism is true, then, surely, so is evolution. Naturalism, therefore, is unacceptable.

The traditional theist, on the other hand, isn't forced into that appalling loop. On this point his set of beliefs is stable. He has no corresponding reason for doubting that it is a purpose of our cognitive systems to produce true beliefs, nor any reason for thinking that P(R/N&E) is low, nor any reason for thinking the probability of a belief's being true, given that it is a product of his cognitive faculties, is no better than in the neighborhood of 1/2. He may indeed endorse some form of evolution; but if he does, it will be a form of evolution guided and orchestrated by God. And *qua* traditional theist — *qua* Jewish, Moslem or Christian theist[20] — he believes that God is the premier knower and has created us human beings in his image, an important part of which involves his endowing them with a reflection of his powers as a knower.[21]

Of course he can't sensibly *argue* that in fact our beliefs are mostly true, from the premise that we have been created by God in his image. More pre-

20 Things may stand differently with a bare theist — one who holds only that there is an omnipotent, omniscient and wholly good creator, but does not add that God has created humankind in his own image.

21 Of course God's knowledge is significantly different from human knowledge: God has not been designed and does not have a design plan (in the sense of that term in which it applies to human beings). When applied to both God and human beings, such terms as 'design plan', 'proper function' and 'knowledge', as Aquinas pointed out, apply analogously rather than univocally. What precisely is the analogy in this case? Multifarious, of course (for example, divine knowledge as well as human knowledge requires both belief and truth); but perhaps the central analogy lies in the following direction. God has not been designed; still, there is a way in which (if I may say so) his cognitive or epistemic faculties work. This way is given by his being essentially omniscient and necessarily existent: God is essentially omniscient, but also a necessary being, so that it is a necessary truth that God believes a proposition A if and only if A is true. Call that way of working 'W'. W is something like an ideal for cognitive beings — beings capable of holding beliefs, seeing connections between propositions, and holding true beliefs. It is an ideal in the following sense. Say that a cognitive design plan P is more excellent than a design plan P* just if a being designed according to P would be epistemically or cognitively more excellent than one designed according to P*. (Of course there will be environmental relativity here; furthermore, one thing that will figure into the comparison between a pair of design plans will be stability of its reliability under change of environment.) Add W to the set to be ordered. Then perhaps the resulting ordering will not be connected; there may be elements that are incomparable. But there will be a maximal element under the ordering: W. W, therefore, is an ideal for cognitive design plans, and it is (partly) in virtue of that relation that the term 'knowledge' is analogically extended to apply to God.

cisely, he can't sensibly follow Descartes, who started from a condition of general doubt about whether our cognitive nature is reliable, and then used his theistic belief as a premise in an argument designed to resolve that doubt. Here Thomas Reid is surely right:

> "Descartes certainly made a false step in this matter, for having suggested this doubt among others — that whatever evidence he might have from his consciousness, his senses, his memory, or his reason, yet possibly some malignant being had given him those faculties on purpose to impose upon him; and therefore, that they are not to be trusted without a proper voucher. To remove this doubt, he endeavours to prove the being of a Deity who is no deceiver; whence he concludes, that the faculties he had given him are true and worthy to be trusted.
>
> It is strange that so acute a reasoner did not perceive that in this reasoning there is evidently a begging of the question.
>
> For, if our faculties be fallacious, why may they not deceive us in this reasoning as well as in others?"[22]

Suppose, therefore, you find yourself with the doubt that our cognitive faculties produce truth: you can't quell that doubt by producing an argument about God and his veracity, or indeed, any argument at all; for the argument, of course, will be under as much suspicion as its source. Here no argument will help you; here salvation will have to be by grace, not by works. But the theist has nothing impelling him in the direction of such skepticism in the first place; no element of his noetic system points in that direction; there are no propositions he already accepts just by way of being a theist, which together with forms of reasoning (the defeater system, for example) lead to the rejection of the belief that our cognitive faculties have the apprehension of truth as their purpose and for the most part fulfill that purpose. On this point, therefore, traditional theism enjoys a clear advantange over the conjunction of naturalism with evolutionary theory.

22 Reid, Th., Essays on the Intellectual Powers of the Human Mind, in: The Works of Thomas Reid. Ed. William Hamilton. Edinburgh: James Thin 1895, Essay VI 447b.