Naturalism Defeated

In the last chapter of *Warrant and Proper Function*[1] I proposed an "evolutionary argument against naturalism". Take philosophical naturalism to be the belief that there aren't any supernatural beings--no such person as God, for example, but also no other supernatural entities.[2] My claim was 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 beams in the edifice of the former.[3] More particularly, I argued 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 argued 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. Now oddly enough, not everyone who has heard this argument has leapt to embrace it; there have been a number of fascinating objections, some published[4] and some unpublished. These objections for the most part revolve around the notion

---

[1] Oxford: 1993 (Hereafter 'WPF').

[2] If my project were giving an analysis of philosophical naturalism, more would have to be said (precisely what, for example, is a supernatural being?); for present purposes we can ignore the niceties.

[3] 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), pp. 6-7.

[4] Among the published and semi-published objections are William Alston's 0000, presented at a conference at Santa Clara University in 19xx, Carl Ginet's 00000, forthcoming in *Synthese*, Timothy O'Connor's "An Evolutionary Argument Against Naturalism?", forthcoming in *The Canadian Journal of Philosophy*, Richard Otte's 0000, presented at the same symposium as Alston's comment, Glenn Ross's 00000 and David Hunt's 0000 presented at the Pacific Division meetings of the APA in March of 1994, Leopold Stubenberg's 00000 presented at a colloquium

of a defeater--a notion crucial to contemporary epistemology, but so far largely unexplored. In this paper I want to examine and respond to those objections, in the process hoping to learn something about defeaters.

I The Argument

Since you may not have a copy of WPF on your desk at the moment, I'll briefly outline the original argument here. It 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 turn out to have survival value and to enhance fitness; they spread through the population and persist. According to this story, 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 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

---

at the University of Notre Dame in 19xx, and 0000's oooo presented at the Central division meetings of the APA in April, 1995.

[5]*Very* roughly: a thermometer stuck on 72 degrees isn't reliable even if it is located somewhere (San Diego?) where it is 72 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.

life, God himself, and much more.[6]   And the above evolutionary account of our origins 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 like 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.   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

---

[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 and arranged by 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 as well supported by the evidence; if so, T is not more probable than not with respect to the relevant evidence.

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?[8] 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.[9]

What Churchland means, I think, is that evolution is interested (so to speak) only in adaptive *behavior*, 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, is best understood as the suggestion that the objective[10] 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

---

[8]Letter to William Graham, Down, July 3rd, 1881. In *The Life and Letters of Charles Darwin Including an Autobiographical Chapter*, ed. Francis Darwin (London: John Murray, Albermarle Street, 1887), Volume 1, pp. 315-316.
[9]*Journal of Philosophy* (LXXXIV, Oct. 87) p. 548.
[10]For an account of objective probability, see WPF, pp. 0000.

true, God might be directing and orchestrating the variation 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.[11]   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? 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 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. 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 most of our beliefs are true, and unlikely that our cognitive faculties are for the most part reliable. 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.[12]

---

[11]In WPF the probability at issue was the slightly more complex P(R/N&E&C), where C was a proposition setting out some of the main features of our cognitive system (see WPF, p. 220). I now think the additional complexity unnecessary.

[12]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

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 suggested 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 direct 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*:[13] their behavior is not caused by their beliefs. On this

Knowledge" *in Kant oder Hegel?* ed. Dieter Henrich (Stuttgart: Klett-Cotta Buchhandlung, 1983) p. 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 may not wish to commit himself. See WPF pp. 80-81.

[13]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 stages of consciousness may, conversely, cause olecular 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." T. H. Huxley "On the Hypothesis that Animals are Automate and its History" (1874), chapter 5 of his *Method and Results* (London, Macmillan, 1893) pp. 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

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. 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 could 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 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

---

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*'.)

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* [14] 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 popular among contemporary philosophers of mind; indeed, Robert Cummins goes so faras to call it the "received view."[15] On this view, as on the last, P(R/N&E) (specified to those creatures) will be 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 belief, then once more beliefs--or rather, their semantic properties, including truth and falsehood--will be invisible to natural selection.[16]

---

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

[15]*Meaning and Mental Representation* (Cambridge, MA: MIT Press, 1989), p. 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.

[16]We must also consider here the possibility that the syntax and semantics of belief are the effects of a common cause: perhapsthere 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

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--'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 indeed both causally connected with their behavior 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? 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

events? But perhaps the right answer is not that the probability in question is low, but that it is inscrutable: see below, pp. 0000.

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 pp. 225-226).

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.

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:

$$P(R/N\&E) = (P(R/N\&E\&P_1) \times P(P_1/N\&E)) + (P(R/N\&E\&P_2) \times P(P_2/N\&E)) +$$

$$(P(R/N\&E\&P_3) \times P(P_3/N\&E)) + (P(R/N\&E\&P_4) \times P(P_4/N\&E)).$$

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. But that is all we need for the argument. For consider the left-hand multiplicand in each of the four terms on the right-hand side of the equation. In the first three, the sensible estimate would put the value low, considerably less that 1/2; in the 4th, it isn't very clear what the value would be, but it couldn't be much more than 1/2. But then (since the probabilities of P1 and of P2 (the two forms of epiphenomenalism) would be fairly high, given naturalism, and since the right hand multiplicands in the four terms cannot sum to more than 1) that means that the value of P(R/N&E) will be less than 1/2; and that is enough for the argument.

But the argument for a low estimate of P(R/N&E) is by no means irresistible; our estimates of the various probabilities involved in estimating P(R/N&E) with respect to that hypothetical population were (naturally enough) both imprecise and poorly grounded. You might reasonably hold, therefore, that the right course here is simple agnosticism: one just doesn't know what that probability is. You doubt that it is very high; but you aren't prepared to say that it is low: you have no definite opinion at all as to what that probability might be. Then this probability is *inscrutable* for you. This too seems a sensible attitude to take. The sensible thing to think, then, is that P(R/N&E) is either low or inscrutable.

Now return to Darwin's doubt, and 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 was 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 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, 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[17] my natural belief that my cognitive faculties are in fact reliable.

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% 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

---

[17]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.

defeater of my original belief: a defeater-defeater as we might put it.[18] But (to return to our argument) can the defeater the naturalist has for R be in turn defeated? I argued that it can't (WPF 233-234). It could be defeated only by something--an argument, for example, that involves some other *belief* (perhaps as premise). But any such belief will be subject to the very same defeater as R is. So this defeater can't be defeated.[19]

But if I have an undefeated defeater for R, 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 he holds--a defeater, as I put it, that is ultimately undefeated But this means, then, that he has an ultimately undefeated defeater for N&E itself. And *that* means that the conjunction of naturalism with evolution is self-defeating, such that one can't rationally accept it.

I went on to add that 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 is apprised of this argument and sees the connections between N&E and R.

II Objections

Now I believe this argument, while inevitably a bit sketchy, has a great deal to be said for it. My exalted opinion of it, however, has not sufficed to protect it from a number of extremely interesting objections. Despite the objections I continue to believe that the argument is a good one: I therefore want to examine and reply to the objections. I have a further and ulterior motive. The objections have to do crucially with the notion of *defeaters*; although this

---

[18]As in fact John Pollock *does* put it.
[19]I'll qualify this below, pp. 0000, when we get to the subject of loops.

notion is absolutely central to contemporary epistemology, it has so far received little by way of concentrated attention.[20]   Exploring these objections will give us, as an added bonus, a good chance to learn something about defeaters. I shall first briefly set out the objections, then make some suggestions as to how defeaters work, and then assess the objections in the light of what (I hope) we will have learned about defeaters.

A. The Perspiration Objection (Michael DePaul, Frederick Suppe, Stephen Wykstra, others). This objection goes as follows. "You claim that the naturalist has a defeater for R in the fact that the probability of R on N&E is either low or inscrutable. But this can't be right. The probability that the function of perspiration is to cool the body, given (just) N&E, is also low, as is the probability that Holland, Michigan is 30 miles from Grand Rapids, given N&E. But surely it would be absurd to claim that these facts give the partisan of N&E a defeater for those beliefs.

B. Austere Theism a defeater for Theism *Simpliciter*? (Earl Conee, Richard Feldman, Theodore Sider, Stephen Wykstra, others) This objection comes in three varieties.

1. "If you are a theist, then, unless your inferential powers are severely limited, you also accept *austere* theism, the view that there exists an extremely powerful and knowledgeable being. But the probability of theism with respect to austere theism, like that of R with respect to N&E, is low or inscrutable; hence (if the principles underlying your argument against N&E are correct) austere theism furnishes the theist with a defeater for theism. But every theist is an austere theist: so every theist has a defeater for theism. Furthermore, this defeater can't be defeated, as is shown by an argument exactly paralleling the one you gave for supposing that the defeater for N&E can't be defeated. So if your argument is correct, the theist has an ultimately undefeated defeater for theism."

---

[20]But see below, footnote 000, for references to the work of John Pollock on defeaters; and see Peter Klein's "Knowledge, Causality, and Defeasibility", *The Journal of Philosophy*, 00000, "Misleading Evidence and the Restoration of Justification" *Philosophical Studies* 37 (1980), and "Immune Belief Systems", *Philosophical Topics*, Vol XIV, No. 1 (1986). Klein is for the most part concerned with "misleading" defeaters.

2. "If N&E is self-defeating in the way you suggest, then so is austere theism. For relative to austere theism, the probability of R is low or inscrutable; the austere theist therefore has an undefeatable defeater for R, but then also for any other belief she holds, including austere theism itself. Austere theism, therefore, is self-referentially self-refuting (if N&E is) and hence cannot rationally be accepted. But of course theism entails austere theism; if it is irrational to accept a proposition *p*, it is also irrational to accept any proposition that entails it; hence if the argument defeats naturalism, it pays the same compliment to theism."

3. "As in (b), the probability of R on austere theism is low or inscrutable; so the theist has a defeater for R, and hence for anything else he believes; but then he has a defeater for theism, one he can't lose as long as he accepts theism."

C. Can't the Naturalist Just Add a Little Something? (Fred Dretske, Carl Ginet, Timothy O'Connor, Richard Otte, John Perry, Ernest Sosa, Stephen Wykstra, others). An austere theist who wasn't also a theist would face the same defeater as the partisan of N&E. Although the theist also accepts austere theism, she escapes defeat because she accepts not just austere theism, but something additional, the difference, we might say, between austere theism and theism. But if it is right and proper for the theist thus to elude defeat, why can't the naturalist do the same thing? Thus Ginet:. . . if we delete this component (the difference between theism *simpliciter* and austere theism) and consider just the hypothesis T- that there is a perfect being who creates everything else, then it looks as if we could argue in just the same way Plantinga argues concerning P(R/N&E&C[21]) to the dismal conclusion that P(R/T-&E&C) is low or unknown. Now how is it that the theist is allowed to build into her metaphysical hypothesis something that entails R or a high probability of R but the naturalist isn't? Why isn't it just as reasonable for the naturalist to take it as one of the tenets of naturalism that our cognitive systems are on the whole reliable (especially since it seems to be in our nature to have it as a basic belief)?[22]

---

[21]See footnote 11.

[22]"000000", *Philosophy and Phenomenological Research*, 000 000, p. 0000. Compare Timothy O'Connor "An Evolutionary Argument Against Naturalism?" *Canadian Journal of Philosophy*: ". . . why can't she [the naturalist] say that her beliefs on these matters are not limited to N&E

D. The Maximal Warrant Objection (William Alston, Timothy O'Connor, William Craig, others)   According to this objection, R has a great deal of intrinsic warrant for us. This proposition has warrant in the basic way: it doesn't get its warrant by way of being accepted on the evidential basis of other propositions. It has so much intrinsic warrant, in fact, that it can't be defeated--or at any rate can't be defeated by the fact that P(R/N&E) is low or inscrutable. A variant of this objection (Van Fraassen) addresses and rejects the argument's implicit premise that if the right attitude towards P(R/N&E) with respect to *that hypothetical population* is low or inscrutable, then the same goes for that probability with respect to *myself* (ourselves).)

E. The Dreaded Loop Objection (Richard Otte, Glenn Ross, David Hunt, others).

Following Hume (and Sextus Empiricus) I said that if the devotee of N (or N&E) is rational, then he will fall into the following sort of diachronic loop: first, he believes N&E and sees that this gives him a defeater for R, and hence for N&E; so then he stops believing N&E; but then he *loses* his defeater for R and N&E; then presumably those beliefs come flooding back; but then once again he has a defeater for them; and so on, round and round the loop. In this loop N&E keeps getting alternately defeated and reprieved--i.e., at t1 it is defeated, at t2 undefeated, at t3 defeated, and so on. And then I went on to say that his falling into this loop gives him an ultimately undefeated defeater for N&E.

According to the objector, the problem is two-fold. First, suppose the devotee of N&E were to fall into such a loop and doggedly plod around it: he wouldn't thereby have an ultimately undefeated defeater for N&E. What he would have instead is a defeater that is not ultimately defeated--a different matter altogether. An ultimately undefeated defeater would be one such that at a certain point it is undefeated, and remains undefeated thereafter. But, says the objector, that doesn't happen here: here the devotee of N&E alternately has and loses his defeater for N&E. For every time at which he has a defeater for N&E, there is a subsequent time at which

---

alone, but include O as well, where O is simply a general proposition to the effect that the initial conditions of the development of organic life and the sum total of evolutionary processes (including ones as yet unknown or only dimly understood) were and are such as to render P(R/N&E&C&O) rather high?" p. 00000.

that defeater is defeated; this alternation is terminated only by death or disability. Hence, obviously, he does not have an ultimately undefeated defeater for N&E.

But further, why think in the first place that rationality requires him to fall into this appalling loop? The fact is rationality requires that he stay out of the loop, or at least get out of it after a couple of tours around it. Can't he see in advance what is coming? He'd have to be (at best) extremely imperceptive to keep on slogging round and round that loop.

III Defeaters and Defeat

A formidable, indeed frightening array of objections: what can be said by way of reply? Note first that these objections all concern the behavior of defeaters in one way or another. Therefore proper procedure demands, I think, that we begin by trying to think to some purpose about defeaters and how they work.

Classical foundationalists such as Descartes had little need for the notion of a defeater: in a well-run epistemic establishment, so he thought,[23] the basic beliefs (those not accepted on the evidential basis of other beliefs) are *certain* and the beliefs in the superstructure follow from those basic beliefs by way of argument forms whose corresponding conditionals are themselves certain. Furthermore, there will be no consequences C1 and C2 of the foundational beliefs such that C1 is improbable, epistemically or objectively, with respect to C2. But then this structure of beliefs will never include a pair of propositions or beliefs one of which is a defeater for the other. With the rejection of Cartesian classical foundationalism, however, defeaters assume a real importance. Locke held that in a healthy structure of beliefs, the relation between the basic beliefs and a belief in the superstructure (a nonfoundational belief) need not be deductive: probability will do. But then of course it could be that a superstructural belief is probable with respect to one element of the foundation but improbable with respect to the conjunction of that element with other elements of the foundation: those other elements, then, will serve as a defeater for the proposition in question. Locke as well as Descartes, however, accepted the

---

[23]Or at any rate as it might initially *seem* he thought: see Nicholas Wolterstorff, 00000

classical foundationalist doctrine according to which properly basic beliefs (those properly accepted in the basic way) are *certain*. But this doctrine is now widely recognized as a snare and a delusion. For example, I now believe in the basic way that I am seated before my computer, that the tiger lilies in the backyard are blooming, and that I had a grapefruit for breakfast: none of these is certain in the Cartesian-Lockean sense; but each is properly basic. And if basic beliefs need not be certain, then there is still another use for the notion of a defeater. For then it can be, not merely that we can acquire defeaters for superstructural beliefs; then we can also acquire defeaters for what we believe in the basic way.

Given the importance of the notion of defeaters for contemporary, post-classical foundationalist epistemology, it is a bit puzzling that this idea has only recently assumed center stage. We find it in Roderick Chisholm's work, first in his "The Ethics of Requirement" *American Philosophical Quarterly* Vol I (164) and second in the first edition of *Theory of Knowledge* (Prentice-Hall, 1966) p. 48; the notion of defeasible reasoning assumes an increasingly large role in the two subsequent editions of that work. Of course this idea of defeasibility goes back long before Chisholm. A *locus classicus* of the notion, and the origin of the use of the terms 'defeat' and 'defeasibility' in this general kind of connection, is H. L. A. Hart's "The Ascription of Responsibility and Rights".[24] The notion of defeat and defeasible reasoning is intimately connected with total evidence and the fact that various strands within one's total evidence can point in different directions; this peculiarity of nondeductive reasoning is noted much before Chisholm by William Kneale,[25] J. M. Keynes,[26] Bernard Bolzano,[27] and

---

[24]*Aristotelian Society, Proceedings*, 1948-49. Hart, however, speaks of *concepts* as being defeasible; he links this with there being no necessary and sufficient conditions for the application of the concept, and links *that*, in turn, with the concept's being such that words expressing it do not denote anything (mental states such as intention and consent, for example) but instead are properly applied on the basis of congeries of criteria: "But the defence, *e.g.*, that B entered into a contract with A as a result of the undue influence exerted upon him by A, is not evidence of the absence of a factor called 'true consent', but one of the multiple criteria for the use of the phrase 'no true consent'" (p. 178).

[25]*Probability and Induction* (Oxford: Clarendon Press, 1949), pp. 9-11.

[26]*A Treatise on Probability* (London: Macmillan and Co. Ltd, 1921) p. 4.

indeed by Locke himself.[28]   But none of these thinkers made any effort to develop and explore this notion.

Now a contemporary cynosure of defeasibility obviously is (or ought to be) John Pollock.[29] And according to Pollock,

(n) If P is a reason for S to believe Q, R is a *defeater* for this reason if and only if R is logically consistent with P and (P&R) is not a reason for S to believe Q.[30]

It looks initially as if a defeater *R* must be a *proposition* or *belief*: presumably it is only propositions or beliefs that stand to something or other in the relationship of being logically consistent. This probably doesn't represent Pollock's intentions, however. For he also believes that an *experience*, sensuous or otherwise, can function as a reason for a belief; but then presumably another experience can function as a defeater for that (experiential) reason. True, (n) wouldn't then be exactly right (given that the defeater is supposed to be *consistent* with P), but it already suffers from the same defect insofar as Pollock thinks experiences can be reasons, and can be subject to defeat. And a proposition *D* is a defeater of a reason *R* for a belief *B*, says Pollock, just if *D* is consistent with *R* and *R&D* is not a reason for *B*.

Many questions arise about this account of defeaters and defeat. Many of the most important revolve about the *nature* of defeaters and defeat. First, what, fundamentally, *are* defeaters? What (if anything) are they for and what do they do? Second, what is this 'is a reason for' relationship that holds between *R* and *B* but fails to hold between (*P&R*) and *B*? Pollock apparently assumes that it is a *broadly logical* relation; at any rate he seems to hold ("Oscar", p.

[27]*Theory of Science*, ed. Rolf George (Berkeley: University of California Press, 1972) pp. 238. (This work was first published in 1837.)

[28]*Essay*, Bk. IV, Ch. xv, xvi sec. 1.

[29]See his "the Structure of Epistemic Justification", *American Philosophical Quarterly*, monograph series 4, (date) p. 62, his *Knowledge and Justification* (Princeton: Princeton University Press, 1974), *Contemporary Theories of Knowledge* (Totowa, NJ: Rowman & Littlefield, 1986) and "The Building of Oscar", *Philosophical Perspectives, 2, Epistemology, 1988*, ed. James Tomberlin (Atascadero: Ridgeview Publishing Co., 1988.

[30]*Contemporary Theories of Knowledge* (Towato, N.J: Rowman & Littlefield, 1986), p. 38. See also "The Building of Oscar" pp. 318-320.

318) that if *R* is a reason for *B*, then it's *necessary* that it is. Is this true? Or might it be that (for some *R*'s and *B*'s) *B* is *in fact* a defeater for *R*, but could have failed to be? Third, if you acquire a defeater (an undefeated defeater) for a belief *B* but continue to hold *B* with unabated strength, what, precisely, is wrong with you? To what sort of criticism are you appropriately subject? Fourth, Pollock holds that a defeater is always a defeater for a *reason* for a belief; is that right, or could a defeater defeat, not a reason for a belief, but the belief itself? Fifth, suppose you have a defeater D for a reason R for a belief B you hold: is (as Pollock seems to hold) D a defeater for B *in itself*, so to speak, or only relative to something else, perhaps something like a body of background information? And sixth, I claim that N&E is self defeating, constitutes a defeater for itself: but is that really possible?

These are perhaps the central questions, but many other important questions arise as well. For example, is it only *beliefs* that can be defeaters, or can other states of an epistemic agent also serve as defeaters? If I hold a belief A and learn something B with respect to which the probability of A is low or inscrutable, does B automatically constitute a defeater (even if perhaps a very easily defeated defeater), for me, of A ? In many central cases, I hold a belief A and then learn something B that is a defeater, for me, of A: but can there also be cases where I already *have* a defeater for A, but perhaps don't realize it? The above evolutionary argument apparently presupposes that a belief can be a defeater *for itself*: but can this really happen? And consider R, the proposition that our cognitive faculties are reliable. Can one really acquire a defeater for R? If so, that defeater would presumably be (or be intimately associated with) some *belief B*; but then if you had a defeater for R, wouldn't you also have a defeater for *B* and hence a defeater-defeater for your defeater for R? Here we seem to verge on a sort of loop--not a diachronic loop, but a loop nonetheless; how are we to think about such loops? We will look into some of these questions, but no more deeply than is necessary for our project.

A. The Nature of Defeat Our question is: what is the nature of defeaters, and what is the nature of defeat? With respect to the first question, the basic but rough answer is that defeaters are reasons for changing one's beliefs in a certain way. An account of defeaters, we might say,

belongs to the subject of *rational kinematics* (apologies to Richard Jeffrey), the subject (if there were such a subject) that specifies the correct or proper ways of changing belief in response to experience and new belief. But what is the nature of defeat? What constitutes correctness or propriety here? Suppose I have a defeater for one of my beliefs *B* and no defeater for that defeater (so that what I have is an undefeated defeater for *B*); but suppose I continue to hold *B* anyway. What, precisely, is my problem? Presumably this is a deplorable state of affairs; even if it isn't a punishable offense, there is something wrong, unhappy, regrettable about it. But *what* precisely, or even approximately?

The usual answer (concurred in, I think, by Pollock) is that I would then be in an *irrational* condition of some kind; there would be something irrational about me, or more precisely about the structure of my beliefs. And this answer, despite its popularity, is correct. But irrationality is multifarious and legion:[31] of what sort are we speaking of here? This is not the place for a taxonomy of rationality (for an initial effort at such a taxonomy, see WCD pp. 000 ff.), but the basic idea, I think, should be something like this. Roughly speaking, a belief, or a withholding of a belief (or a decision, inclination, act of will or other bit of cognitive functioning) is *rational* (in the relevant sense), in a set of circumstances, when it is one that in those circumstances could be displayed or undergone by a rational person. More specifically, what counts is the sort of response that could be displayed or undergone by a rational *human* person; significantly different outputs might be compatible with proper function for angels, or Alpha Centaurians, or other rational creatures with design plans somewhat different from ours. And a rational human being is one whose rational faculties, (or *ratio*, or cognitive faculties) are *functioning properly*, subject to no dysfunction or malfunction. More specifically, it is the faculties or belief producing processes involved in the production of the belief in question that must be functioning properly; the rationality of my belief that China is a large country is not compromised by the fact that I harbor irrational beliefs about my neighbor's dog. What is

---

[31]See my *Warrant : the Current Debate* (New York: Oxford University Press, 1993) pp. 000 (hereafter 'WCD').

relevant, here, is not *ideal* function; we aren't thinking of the way in which an ideally rational person would function. (An ideally rational person, I suppose, would be omniscient, and perhaps a *really* ideally rational person would be essentially omniscient, omniscient in every possible world in which he exists.) What is relevant here is the sort of function displayed by a cognitively healthy human being.[32] Even this leaves some latitude, of course, but for now I think we can live with it.

A belief is rational in a certain set of circumstances and rational when it is a healthy or sane belief to hold in those circumstances. The relevant circumstances have a two-tiered character. First, there is my *noetic structure*: an assemblage of beliefs and experiences (and other cognitive states such as doubts, fears and the like) together with various salient properties of these states and relevant relations obtaining among them. Let's oversimplify and think just of beliefs and experiences. A description of a noetic structure would include a description of the strength of each belief, of the logical relations between that belief and others, and of the circumstances (crucially including experiences) under which the belief in question was formed and sustained. Not all beliefs are formed in response to experience (together with previous belief), and it may be that some beliefs are formed in response to experience, previous belief and still other circumstances: let's use the ugly but popular term *doxastic input* to denote whatever it is that beliefs are formed in response to. Then we can say that a description of *S*'s noetic structure would include an account of the doxastic input to which S has been subject, as well as an account of the doxastic responses thereto. Again, much more should be said, but perhaps for present purposes we can leave the notion of a noetic structure at this intuitive level. We can note parenthetically that a noetic structure can be *rational*; a rational (human) noetic structure is one such that its doxastic output could be the output (for given doxastic input) on the part of a rational human being--one, that is, whose rational faculties have all along been functioning

---

[32]And this enables us to answer another of the questions on p. 000; since what constitutes proper function on the part of human beings could have been at least a bit different, it is not the case that if *A* is a defeater of *B*, then it is *necessary* that it is.

properly. (No belief in a rational noetic structure results from and is sustained by an irrational (insane, dysfunctional) response to circumstances and experience.

The first tier of my current circumstances, therefore, is my noetic structure; and the second tier of circumstances is current experience--more broadly, current doxastic input. And we can say that a belief is rational, with respect to a given noetic structure and given current doxastic input, just if it is a rational (i.e., nonpathological, nondysfunctional) doxastic response to that structure and doxastic input. Note that it is entirely possible for a belief to be rational in this sense even if the noetic structure of which it is a part is not rational. Again, perhaps I hold irrational beliefs about my neighbor's dog (I have finally snapped under the strain of that constant barking and have come to the irrational belief that this dog is purposely trying to drive me insane); some of my beliefs may nonetheless be rational with respect to my noetic structure and current doxastic input. In fact other beliefs about that dog could have that property: for example, the belief that it is now digging up my newly planted lilacs, or even the belief that it would be a good thing to engage a hitman (a canine specialist) from Chicago to take care of it.[33]

By way of example of an irrational, pathological doxastic response: I have never met you, although you have been pointed out to me at a distance, and I have the impression that your name is Sam. We meet at a party, and you tell me your name is not Sam but George; and suppose my noetic structure is more or less standard, at least with respect to experience with people's telling me their names. If, under these conditions, I continue to believe your name is Sam, then there is something wrong with me; my cognitive reaction is not the right or appropriate or normal (in the nonstatistical sense) reaction (given my noetic structure) to that

---

[33]The point here is that *given* a certain (perhaps irrational) belief and circumstances. The rational belief to form is p. The structure here is like that with respect to morality: *given* that you are going to do a certain (perhaps immoral) action A, the moral thing to do is B. Given that you are going to insult me or maim me, the moral thing to do is insult me. Similarly here: given that I have come to believe (perhaps irrationally) that this dog is trying to drive me insane, the rational thing to do is give up my previous view that dogs never intentionally set out to drive people insane.

doxastic input. The same goes more dramatically if I form the belief that you are only 4 inches tall, or that you are an alien from outer space sent to kidnap me.

The rationality involved here is in the same neighborhood as that involved in warrant: that property or quality, or better, quantity, enough of which is what distinguishes knowledge from mere true belief. There isn't space here to explain how rationality as proper function is involved in warrant;[34] roughly, however, a belief has warrant (as I see it) for a person in case that belief is produced by cognitive faculties working properly in an appropriate cognitive environment according to a design plan successfully aimed at the production of true belief. The relation between warrant and defeaters is complex. First, it is not the case that a defeater for one of my beliefs defeats the *warrant* that belief has for me; for I can have a defeater for a belief that has no warrant for me. In WPF, I argued that if I lie to you and you believe me, then the belief you acquire has little if any warrant, even though your cognitive faculties are functioning just as they should. The reason is that warrant requires more than just that *your* faculties be functioning properly (see WPF, pp. 000 ff.); the rest of your cognitive situation, including your cognitive environment, must also meet certain conditions. But of course you can acquire a defeater for a belief you have as a result of my lying to you (perhaps I shamefacedly confess the lie); so you can acquire a defeater for a belief that has little or no warrant for you. What is important is that your new belief (that I lied) rationally requires a revision in your structure of beliefs. On the other hand, if I acquire an undefeated defeater for a belief but continue to hold it, that belief will not have warrant for me.

Furthermore, a circumstance that defeats the warrant a belief has for me need not be a defeater for that belief: it need not make it irrational for me to continue to hold the belief. Unbeknownst to me, you build a lot of fake barns in my neighborhood; I see what is in fact a real barn and form the belief that it is indeed a real barn; all those fake barns prevent this belief from having much by way of warrant for me, but they (or the proposition that they are present around

---

[34]See WPF chaps 1 and 2.

here) do not constitute a defeater, for me, for that belief. (Of course if you had *told* me you had done this deed, then I would have had a defeater for the belief in question.)[35] Still further, a belief *D* can defeat another belief *B*, for me, even if *D* has little or no warrant for me. I believe that there is a sheep in the pasture; you (the owner of the meadow) tell me there are no sheep in the neighborhood, although (as you add) you do own a dog who frequents the pasture and is indistinguishable from a sheep at 100 yards; I believe you. As it happens, you are lying (in order to demonstrate your liberation from pre-post-modern ideas about truth and truth-telling). The belief I innocently form as a result of your mendacity has little or no warrant for me but is nonetheless a defeater for my belief that there is a sheep in the pasture.

This example takes advantage of the fact that warrant sometimes requires more than proper function on the part of the believer; but other examples show that it is also possible for a belief formed by way of cognitive malfunction to serve as a defeater. Years of giving in to my unduly suspicious nature may finally lead me to think my best friend is secretly out to get me; that belief is irrational, has no warrant for me, but can nonetheless serve as a defeater for my belief that best friends never do things like this. *Given* that I believe my friend is out to get me, it isn't rational to believe that nobody's friend is ever out to get him, even if my belief that my friend is out to get me is itself irrationally formed. (If you like more fanciful examples, note that if I am a brain in a vat, or deceived by a Cartesian demon, or am struck by an errant burst of cosmic radiation (causing me to believe that I have just won the Nobel prize in chemistry) the beliefs I form as a result of these malfunctions can function as defeaters, for me, even if they lack warrant.)   Finally, what rationality requires, in the case of a defeated belief, is that I *withhold* that belief, fail to accept it. Withholding a belief, of course, won't have warrant: it won't have the

---

[35]Contrast Peter Klein, 000000. Klein's conception of a defeater, so far as I can see, is of something that defeats the *warrant* a belief enjoys. This is a perfectly sensible way to think about defeaters: more exactly, it is perfectly sensible to think that there *are* defeaters of this sort. My concern, however, is with defeaters that defeat the *rationality* of a belief, not its warrant. Both kinds of defeaters are important; rationality defeaters are what are relevant in this context, in which we are thinking about the rationality of a certain belief (N&E), not about its warrant.

property that distinguishes knowledge from mere true belief. Still, the rationality involved is of the same kind: the rational reaction(s), here, will be one(s) dictated or permitted by the human design plan (in the circumstances that obtain); it will be the sort that could be displayed by a person with properly functioning cognitive faculties in those circumstances.

B. Do Defeaters Defeat *Reasons*? According to Pollock, a defeater is a defeater for a *reason* for a proposition. That reason, he thinks, may be another proposition or belief, but it needn't be; it could also be an experience, for example a way of being appeared to. Return to the sheep in the pasture: my original reason for thinking that there is a sheep there is that it *looks* to me as if there is a sheep there. (We might say, with apologies to Roderick Chisholm, that I am appeared to sheeply). When you tell me the lie, I acquire a defeater; my defeated reason is not a proposition or belief, however, but instead a visual experience or way of being appeared to. So the defeated reason need not be a belief. But in any case, says Pollock, it is a *reason* that gets defeated, not the proposition or belief for which that reason is a reason. So he says; but is he right here?

In the argument against naturalism, I spoke of a defeater for a *proposition*, a *belief*, a defeater for N, rather than for a reason for N. This is more satisfactory, I think, for a couple of reasons.

First, it is more natural. A defeater (thought of my way) is in essence a reason for withholding a belief; but obviously there are reasons for withholding a belief that do not consist in defeat of reasons for that belief.[36] But more than naturalness is at issue. In some cases you have an apparent defeater for a belief, and there simply *isn't* anything that can be identified as your reason for that belief--no other belief or proposition, surely, but also no way of being appeared to or other experience. Consider memory, for example. I remember that I mowed the lawn yesterday; but what is my reason for so thinking? Not some other belief or proposition--I don't ordinarily infer it from, e.g., the fact that the lawn looks as if it were mowed just yesterday, and I'm the person who mows it. (True, as one ages it may happen more often that this is how

---

[36]So perhaps the fact is Pollock is really concerned with warrant defeaters, not rationality defeaters.

you form beliefs about what you did yesterday; but it's not the usual run of things.) Nor is there some experience--certainly no way of being appeared to--that is my reason for believing I mowed yesterday. My memory that I mowed the lawn may indeed be accompanied by bits of sensuous imagery as of someone mowing a lawn; but that imagery is partial, fragmented, fleeting, hard to focus. It isn't nearly detailed and specific enough to enable me to determine, on the basis of it, that it was *my* lawn that got mowed, that it was *I* who mowed it, and that it was *yesterday* when it all happened.[37] Perhaps this sort of imagery accompanies memory for most of us; nevertheless it isn't a reason for the memory belief formed.

Of course there is also *another* kind of phenomenology involved with memory, a phenomenology that accompanies all beliefs. There is a sense of the belief's being *right*--its *feeling*, somehow, like the correct belief here, its feeling *natural* or proper, or appropriate in the circumstances. The thought that it was *you* (not I) who mowed my lawn yesterday feels strange, alien, unacceptable--it's hard to find the right words. (Where are the phenomenologists, now that we need them?) But this sense or feeling, while it seems to be an experience of some kind, and while it ordinarily accompanies memory beliefs, can hardly be my *reason* for accepting the belief; such a feeling of rightness or propriety accompanies *every* belief, including those that we rightly describe as ones for which I have no reason at all.

So memory beliefs are ones I quite properly accept, but don't accept on the basis of a reason, experiential or otherwise. Of course it doesn't follow that they simply pop into one's mind at random. They are instead occasioned by circumstances. For example, you ask me what I had for breakfast; I reflect for a moment, and the answer comes to mind: oatmeal and an orange.   Your question then (or perhaps the experience involved in my hearing and understanding your question) occasions my memory belief; it is not, of course, my reason for accepting that belief.

---

[37]See WPF, pp. 58-64.

Elementary *a priori* beliefs are also beliefs I don't accept on the basis of reasons. I believe that 3+1 = 4, the corresponding conditional of *modus ponens*, and that no dogs are sets. Of course I don't accept these beliefs on the evidential basis of other beliefs; and while there may be a sort of visual imagery involved (perhaps something resembling a glimpse of fragments of a sentence on a blackboard, or perhaps something like a quick look at a sort of indistinct dog between braces), I don't form the belief on the basis of that imagery.[38] Memory and some *a priori* beliefs, then, are not beliefs that we hold on the basis of reasons. But we do sometimes have defeaters for memory and *a priori* beliefs; such defeaters, therefore, are not defeaters for reasons for such beliefs.

Still further, with respect to some sorts of belief, what gives me a defeater for a belief of that sort is just the fact that I realize that I don't have a reason for it; that realization is itself a defeater for the belief. Due to a small cognitive glitch, for example, the thought that you are now in San Francisco unaccountably pops into my mind; I suddenly form the belief that that's where you are. I then realize that no one has told me that you are there; I have no track record of telepathy or anything of the sort; I have no reason of any sort to think that you are there. But this belief is of the sort requiring reasons, if it is to be accepted rationally; my realizing that I have no reasons is itself a reason, for me, to reject the belief, to be agnostic with respect to that proposition. This realization, then, is a defeater for me for this belief; but of course it is not a defeater for a *reason* for the belief.

C. Defeat is relative A defeater of a proposition, whether potential or actual, is always, *pace* Pollock, a defeater *with respect to* a given noetic structure. You and I both believe that the University of Aberdeen was founded in 1495; you but not I know that the current guidebook to Aberdeen contains an egregious error on this matter. We both win a copy of the guidebook in the Scottish national lottery; we both read it; sadly enough it contains the wholly mistaken affirmation that the university was founded in 1595. I thereby acquire a defeater for my belief

---

[38]For more on the phenomenology of memory and *a priori* beliefs and reasons, see WPF, 0000.

that the university was founded in 1495, but you, knowing about this notorious error, do not. The difference, of course, is with respect to the rest of what we know or believe: given the rest of what *I* believe, I now have a reason to reject the belief that the university was founded in 1495, but the same does not hold for you. You already know that the current guidebook contains an egregious error on the matter of the date of the university's foundation; this neutralizes in advance (as we might put it) the defeating potential of the newly acquired bit of knowledge, vis., that the current guidebook to Aberdeen says the university was founded in 1595. Your learning that the guidebook gives 1595 as the date of the university's founding does not give you a defeater for your belief that it was founded in 1495. Hence this new bit of knowledge is a defeater, for me, for the belief that the university was founded in 1495, but not for you. The reason, obviously, is that it is a defeater for that belief with respect to my noetic structure but not with respect to yours. Here we might also note that a belief *D* can serve as a defeater for a belief *B*, but would not have served as a defeater for the conjunction of *B* with some other proposition; to put it a bit misleadingly, *D* can be a defeater of a belief, but fail to defeat the conjunction of that belief with another belief. As before, my belief D that the guidebook says the university was founded in 1595 is a defeater for my belief that the university was founded in 1495; but if instead I had believed both that the university was founded in 1495 and that the guidebook contains an egregious error about when it was founded, D would not have been a defeater for that conjunction (or for either conjunct).

D. How defeaters work We now have answers to the first group of questions on p. 00023. It is time to attempt a characterization of defeat and the defeater relation; but first, suppose we consider a couple of paradigm cases of defeaters. Recall the problem with the current guidebook to Aberdeen: I visit Aberdeen, read the guidebook, and come to acquire the mistaken belief that the university was established in 1595. But then I attend a local reading of the poetry of William McGonagall, poet and tragedian; in the course of the proceedings someone mentions the mistake and the mortified author of the guidebook stands up and acknowledges his grievous error. I then acquire a defeater for my previous belief, a reason for withholding it, a reason (in this case) for

believing something incompatible with it. If by some odd chance I continue to believe that the university was established in 1595, then by that token I would be displaying irrationality. Another example we have already encountered: I see what looks like a sheep in the adjoining pasture; you, the owner of the field, then tell me that you keep no sheep there, but do own a dog visually indistinguishable from a sheep at moderate distances. Again, if circumstances are more or less standard (for example, I have no reason to think you are lying) and I continue to believe that the pasture contains a sheep, I would be displaying irrationality, cognitive dysfunction. Still another example, this one due to John Pollock: I enter a factory and observe an assembly line on which there are widgets spaced at 15 inch intervals; they look red, and I form the belief that they are red. But then the shop superintendent happens along and tells me that the widgets are irradiated with infrared light, making it possible to detect otherwise undetectable hairline cracks. Realizing that those widgets would look red no matter what their color, I have a defeater for my belief that the widgets are red. In this case, unlike the previous two cases, I don't acquire a reason for believing that the widgets are *not* red (as I would if the shop superintendent added that they were really white); it is rather that I lose my reason for thinking that they *are* red. That reason is canceled by the defeater. I shall follow Pollock in calling defeaters of this kind 'undercutting' defeaters; defeaters that give me a reason for believing the denial of the proposition defeated are 'rebutting' defeaters.

In this simplest sort of case, *S* holds a belief *B*, but then she learns something new, *D*; and the rational response to learning *D* is to withhold *B* (or believe it less firmly). But this is only the simplest case. Here is one kind of complication: we want to be able to consider the sort of case where it seems that at a given time *t* you already have a defeater for a belief, but don't realize it: you don't see the relevance of the one belief to the other; you don't make the connection. You are a detective investigating a murder. You think the butler is innocent (he has such a wholesome look about him), but you also know facts f1 - fn about the case, which taken together imply that the butler must after all be the guilty party (no one else had opportunity). Should we say that you really had a defeater all along, since you knew f1 - fn all along, and with respect to

them it is at best extremely unlikely that the butler is innocent? Or should we say that you didn't really have a defeater for this belief until you *realized* that the butler's innocence is unlikely, given f1 - fn? Suppose we file this question for future reference; perhaps our account of defeat will enable us to see an easy answer.

Another kind of complication: perhaps the new thing I learn doesn't require that I withhold a belief B I have been accepting; perhaps it only requires that I hold it less strongly. For example, I believe very firmly that your name is 'Mathilda' and have believed this for some time. But then you tell me that it really isn't 'Mathilda', but 'Letitia'; you have never liked the latter, however, and hence tell everyone that your name is 'Mathilda'. But your manner, in telling me this latter, is ironic and a bit enigmatic; furthermore you have often told me outrageous things with a straight face. I'm *inclined* to think you are only joking, but I can't be really sure. Then, I should think, the rational degree of belief, for me, in the proposition that your name is 'Mathilda' is less than it was. If I continue to accept it with the same fervor as before, there is something irrational about me or my noetic structure.

A further matter here: must a defeater be a *belief*, or could some other cognitive condition serve as well? A defeat*ee* will be a belief--the belief that *P*, for some proposition *P*. And often what defeats this belief, if it gets defeated, will be some other belief. But not necessarily always. I tell you that there are no tulips in Holland, Michigan this May: too much cold weather in April. You are obliged to go to Holland to visit a sick aunt; as it happens, you go in May. Driving into the city, you are confronted by a splendid field of tulips in full bloom. You then have a defeater for your belief that there are no tulips there then, whether or not you explicitly form the belief-- "Hey! There are *lots* of tulips here." So perhaps defeaters need not always be beliefs. Even when they are not, however, there will be a belief relevantly associated with the defeater. In this case, for example, the defeater is an experience, a being appeared to in a certain way; that experience is such that, given your noetic structure, it would be rational to form the belief that there are tulips in Holland in response to it; and if you *had* formed that belief, it would have been

a defeater for the belief for which the experience is a defeater.   But the relevant associated belief need not actually be formed in order for there to be a defeater.

As we have seen, the basic idea is that when *S* acquires a defeater for *B*, she acquires a reason for modifying her noetic structure in a certain way. Or rather, since putting it that way suggests that this modification is something she voluntarily does, when she acquires a defeater for a belief, then if her cognitive faculties are functioning properly, further change in her noetic structure will occur; rationality (in the sense of proper function) requires a change in the rest of her noetic structure. Of course whenever you acquire a new belief *B* and your cognitive faculties are functioning properly, there will be other changes: no doubt rationality will also require adding some conjunctions, disjunctions and conditionals involving *B*, as well as simple logical consequences of *B*, and perhaps of *B* together with other things you believe. (Or perhaps what is added is not these explicit beliefs, but a disposition to form them upon considering them.) Perhaps certain counterfactuals will be required or permitted, and perhaps still other beliefs you hold will be held more strongly.

It is worth noting that in some cases your can't acquire a new belief without the occurrence of changes to aspects of your noetic structure other than your beliefs--nondoxastic aspects, as we may call them. No doubt your present noetic structure *N* does not contain the belief that there is a small lake named "Damfino" north of Mt. Baker in the North Cascades National Park; you have never heard of any such lake.[39] A noetic structure *N** including the belief that there is such a lake and such that rationality permits your moving to it from *N*, will not differ from *N* just by virtue of containing that belief together with appropriate modifications to other beliefs. The reason is that (apart from special circumstances) you won't rationally acquire this belief without hearing or reading that there is such a lake; and that will occur only if you have the experiences involved in hearing or reading or otherwise learning that there is.

---

[39]As a matter of fact, there are Damfino lakes and creeks and valleys in most of the mountainous areas of the US; according to traditional lore, each received its name when one prospector asked another "Well, where the hell are we *now*?"

Given all this, how shall we state the matter? Many refinements and qualifications will no doubt be necessary, but to a zeroeth approximation,

> (D) *D* is a defeater of *B* for S iff *S'*'s noetic structure *N* includes *B* and is such that any human being (1) whose cognitive faculties are functioning properly in the relevant respects, (2) whose noetic structure is *N*, and (3) who comes to believe *D* but nothing else independent of or stronger than *D* would withhold *B* (or believe it less strongly[40]).

A rebutting defeater, to use Pollock's term, is one that works by giving you evidence against *B* (the evidence might but need not be strong enough to require the belief that not-*B*); an undercutting defeater attacks your reasons, whatever they are, for believing *B*.

We could also put it in terms of the human design plan: given noetic structure *N* and new belief *D*, that design plan requires the deletion of *B* from S's noetic structure. We could also put it like this: *D* is a defeater of *B* for you if your noetic structure includes *B* at *t*; at *t* you come to believe *D*; and rationality requires that if you continue to believe *D*, you will cease believing *B*. Note that on this account of defeat, it is possible to have a defeater for a belief even if your noetic structure is in some respects irrational. This is fairly obvious if the locus of the irrationality, so to speak, is distant from the operation of the defeater. Years of giving in to my tendencies towards self-centeredness and a desire for fame finally take their toll: I now foolishly believe that I will be named one of the ten best dressed men of the year. My noetic structure is therefore irrational to some degree; but this doesn't prevent me from having a defeater for my belief, say, that you have just inherited a small fortune. But it is also possible to have a defeater for a belief that is in the same noetic neighborhood as a locus of irrationality. My belief that my neighbor's dog is intentionally trying to annoy me may be formed by way of cognitive

---

[40]We could use the term 'partial defeater' for defeaters that require believing *B* less firmly (as in the Mathilda case above). A full treatment would explain degrees of belief (which are not to be thought of as probability judgments; see WPF pp. 0000) and show how partial and full defeat are related. Here there is no space for that: but note that defeat is really a special case of partial defeat, at least if we stipulate that coming to withhold *B* is a special case of coming to believe *B* less strongly. For the sake of brevity I will henceforth suppress mention of partial defeaters, although the application of what I say to them should be routine.

malfunction and thus irrational; it can nevertheless function as a defeater for a previously held belief that dogs never intentionally try to annoy people. In such cases, one's noetic structure prior to the acquisition of the defeater is irrational; if after the acquisition of the defeater, the defeated belief persists, one's noetic structure, we might say, will be *more* irrational. Granted, this notion of degrees of rationality may be a bit dicey; but if we accept it, then we can also put the matter like this:

> (D*) *D* is a defeater for *B* for *S* at *t* if and only if (1) *S* acquires the belief *D* at *t*, (2) *S'*s noetic structure *N* at *t* is such that adding *D* to it results in a noetic structure that is more irrational than *N*, and (3) no noetic structure to which *S* can rationally move (given that she accepts *D*) will contain *B*. (If you prefer the idea of *loci* of irrationality to that of *degrees* of irrationality, we can restate (D*) accordingly.)

Roughly, the idea is that *S*'s noetic structure *N* is such that her learning *D* requires a change in belief: the noetic structure that results from adding *D* to *N* is irrational, and more irrational than *N*. For example, I have thought for some time that you once spent a year in Aberdeen, Scotland; you tell me (soberly, with no hint of teasing or joking) that you have never been to Scotland, although you once planned to go but were prevented at the last moment. I thus learn that you have never been to Scotland; the noetic structure that results from adding this belief to the noetic structure I have at *t* is irrational, in that it would involve my believing both that you have been to Aberdeen and that you have never been to Aberdeen. What rationality requires is that I change belief; it requires is that I not hold just the beliefs I held at *t* with the addition of the belief that you have never been to Aberdeen. The second clause, therefore, says that some change is required.

But not just any old change; the third says that a change is required with respect to **B**. What needs to be given up is my belief that you once visited Aberdeen; it is *that* belief that must be expunged. You might think that other changes would be possible, i.e., consistent with rationality. Perhaps I could give up the belief that you are now truthful, or the belief that you are mentally competent, or the belief that you are capable of distinguishing Aberdeen, Scotland,

from Aberdeen, South Dakota. And perhaps these changes would be rational with respect to *some* noetic structures--ones, perhaps, in which I have enormously powerful evidence for your having been in Aberdeen--newspaper stories complete with pictures, for example--or structures in which I have good reason to doubt that you are telling the truth. But these are quite different from the noetic structure I do in fact display at *t,* which involves my being quite properly sure that you are telling the truth, and also involves my having little more by way of support for the belief that you have been to Scotland, than a sort of vague memory to the effect that I once learned this. With respect to *this* noetic structure, these changes would not be rational. That is why the defeater is a defeater for my belief that you have been in Aberdeen, rather than for some other belief.

Of course a defeater *could* be a defeater for a conjunction (or for a pair of beliefs, whether or not I have considered and explicitly believe their conjunction) without being a defeater for either conjunct; perhaps rationality requires that I give up the conjunction, but is silent on the question which conjunct to give up. Is it possible that a defeater be a defeater for my whole noetic structure, without being a defeater with respect to any smaller set of beliefs? I don't think so; at any rate the most plausible candidate for this post, the so-called 'paradox of the preface' does not fill that bill. Suppose I notice the fact that at most times in the past I have clearly held at least some false beliefs and conclude that no doubt the same is true now: I hold at least one false belief. Then I believe with respect to some set S of my beliefs that at least one member of S is false; but then the total set S* of my beliefs now is inconsistent in the sense that there is no possible world in which each member of it is true. Does this belief that at least one of my beliefs is false give me a defeater for my entire noetic structure without giving me a defeater for any member of that structure? No: for if at *t* I believe that all of my beliefs are true, then (when I see how likely it is that some of my beliefs are false) I acquire a defeater for *that* belief; but if at *t* I do not believe that proposition (or any other that entails it) then rationality need not require a change. It is perfectly rational to hold a set *S* of beliefs while recognizing that no doubt

at least one member of S is false; rationality does not require one's going on to give up some of the members of S.

We must add an account of how it is that a defeater can itself be defeated. To return to John Pollock's example, you come to the factory, see those apparently red widgets on the assembly line and form the belief (1) those widgets are red. But then the shop superintendent tells you that in fact the widgets are irradiated by infrared light, so that they would look red no matter what their color; you now have a defeater for (1), and you no longer believe it. But then the president of the firm comes along and tells you that the shop superintendent, while reliable on most topics, has a thing about widgets and infrared light: he tells everyone this same story, although as a matter of fact the widgets in this factory are never irradiated by red light; the other employees crowd around and confirm the president's testimony, and you believe him. You now have a defeater for the defeater of (1), a defeater-defeater, we might say. It would be possible, of course, that you acquire a defeater for this defeater-defeater (a (defeater-defeater)-defeater), and so on; but suppose we just stick with defeater-defeaters, leaving the further members of this series to their own devices.

There are two simplest cases. First, it might be that you acquire a defeater in the sense of (D) or (D*) for the defeater D itself. That is what happens in the above widget case: you get a defeater for the belief that the widgets are being irradiated by red light and would look red no matter what their color. This defeater-defeater can be either a rebutting or undercutting defeater. (In the above widget case you have a rebutting defeater; you would have had instead an undercutting defeater if you had learned that the shop superintendent was speaking another language, one in which 'The widgets are being irradiated with red light' means that his name is Sam and he is from Arizona.) Call this kind of defeater-defeater an *intrinsic* defeater-defeater. But there is another sort of case as well, one in which you don't get a defeater in the sense of (D) for D, but (if I may put it so) the defeating potential of D is nevertheless neutralized. What happens is that at t your noetic structure N includes B; then at t^* you come to believe something D which is a defeater of B, so that you move to a noetic structure that includes D but not B; then

at t^{**} you learn (come to believe) something D^* such that its addition to your noetic structure permits a move to a noetic structure that includes D, D^* and B. Call a defeater-defeater of this sort a *neutralizing* defeater-defeater. For another example, modify the widget case as follows: imagine that after speaking with the shop superintendent you pick up one of those widgets and take it over to the window, where you can view it in clear daylight; it still looks red. Then you will quite properly believe both that it was irradiated by red light, that when viewed in clear daylight it looks read, and that it is red. Another sort of neutralizing defeater-defeater: I have heard somewhere that you can't swim and at t believe that; at t^* I learn that you are a lifeguard, which (together with my belief that nearly all lifeguards can swim) gives me a defeater for my belief that you can't swim; but then at t^{**} I learn that you are a Frisian lifeguard and that only half of the Frisian lifeguards can swim, which gives me a defeater for that defeater; but then at t^{***} I learn that you graduated from the famous lifeguarding school at Leeuwarden, all of whose graduates can swim, which gives me a defeater for that defeater-defeater; and so on; we can add to the series *ad libitum*. If the last member of the series is odd numbered, I will wind up rationally holding the original belief along with its defeater, the defeater for that defeater, and so on.

Here we can conveniently answer two more of the questions raised on p. 000. First: suppose for some time I have known or believed a proposition P that implies that some other belief B of mine is false (or with respect to which B is unlikely): have I then had a defeater for B all along, so to speak? Or must it also be the case that I see the relation between B and P? Here there are really two questions: (a) can it be that I have in fact had a defeater for one of my beliefs for some time, failing to make the changes required by rationality? The answer, I should think, is yes indeed: but of course only on pain of irrationality. The second question (b) is the more interesting: in order to have a defeater D for one of my beliefs B, must I see the relevant connection or relation between D and B? Ordinarily, I think, the answer is that you do have to

see the relation to have a defeater. Consider Frege before he received Russell's devastating letter.[41] He believed

> (2) For every property or condition, there exists the set of just those things that have the property or display the condition.

Presumably he also believed then that

> (3) There is such a property or condition as being nonselfmembered;

at any rate let's assume that he did. What he didn't see, and what Russell pointed out to him, is the logical relation between these propositions: together they imply that there is such a thing as the set of nonselfmembered sets, and it both is and isn't a member of itself. So (before Russell's letter) did Frege's belief (3) constitute a defeater for his belief (2)?

I think not. Consider Frege's noetic structure N (before Russell's letter) and consider N-(3), a structure appropriately similar to N given that it lacks (3). (Perhaps this would have been Frege's noetic structure then had it never occurred to him to think about the question whether there is such a property.) Surely Frege wasn't irrational (in the sense of displaying cognitive malfunction) in believing both (2) and (3), even if together they (along with some obvious necessary truths) entail a contradiction. You aren't automatically irrational in believing both (2) and (3): the connection between the two is obscure and difficult and it requires a great deal of logical acumen, more than most of us can muster, to discover it. (Of course it requires much less acumen to see the relation once it is pointed out.) So we should say that Frege's defeater for (2) was not just (3), but (3) together with the newly acquired belief that (4) (2) and (3) together entail a contradiction. Although rationality doesn't require seeing the connection between (2) and (3), failure to realize that (3) and (4) together imply the falsehood of (2) would of course be irrational, pathological.[42] But failure to see that (3) by itself does so is not. The same goes for

---

[41]The letter (date ?) in which Russell pointed out that Frege's axioms for naive set theory yielded a contradiction.

[42]Or does someone who comes to see the logical relation between (2) and (3) acquire a defeater for *(3)*? Or a defeater for the conjunction of (2) with (3) and a partial defeater for each of them? We don't have the space here to enter this very interesting question.

the detective of a few pages back: he believed that the butler was innocent, but also knew those facts f1-fn implying that the butler is the only one who could have committed the crime. He didn't have a defeater all along for the belief that the butler is innocent; he acquired such a defeater only upon reflecting upon f1-fn and seeing their bearing on the belief in question. And his defeater for the belief that the butler is innocent is not just the conjunction of those facts f1-fn; it is the latter together with the belief that those facts imply that the butler did it. A more general consequence is that you never acquire a defeater for a belief, without also acquiring a new belief.

And second, consider the question whether a proposition can really be a defeater of itself. Well, why not? Consider the sort of skeptic who holds that his cognitive faculties are unreliable: he believes -R. But then suppose he sees that if -R is true, then all of his beliefs have been formed unreliably, i.e., formed by unreliable cognitive faculties, and suppose he sees that this implies that -R itself is thus unreliably formed. Then he has a defeater for -R, a reason to withhold it; for a rational person who comes to see that one of her beliefs is unreliably formed will indeed withhold that belief.

Given this quick and sketchy tour of defeaters and defeat, we are now ready to return to the objections to the evolutionary argument against naturalism.

IV Replies

A. The Perspiration Objection The first objection, you recall, was that it was improbable, with respect to N&E, that the function of perspiration is to cool the body, or that Holland, Michigan is about 30 miles from Grand Rapids; hence on the principles underlying the evolutionary argument against naturalism I gave, N&E is a defeater for those two beliefs; but surely it isn't; so there is something seriously amiss with principles underlying the argument, and hence with the argument.

But now the answer is clear. First (and parenthetically), of course, if the evolutionary argument is correct, N&E (together with the argument I gave) *is* a defeater for those propositions. The idea was that N&E is a defeater for R, but then consequently also a defeater

for *any* belief held by the partisan of N&E, including beliefs about perspiration and Holland, Michigan. But suppose we take the objection in the spirit in which it is offered: the thought is that on the principles underlying the argument, N&E would be *immediately* or *directly* a defeater for the propositions in question, just because the latter are improbable with respect to the former. Is the objector right? I think not. She apparently presupposes that (or apparently presupposes that my argument presupposes that)

> (5) For any propositions *A* and *B* I believe, if *B* is improbable or inscrutable with respect to *A* (i.e., the right attitude towards the question of its probability with respect to *A* is agnosticism) then *A* is a defeater for *B*.

(5), however, is false, which ill befits a principle. For example, I believe

> (6) You own an old Nissan,

and I also believe

> (7) You own a Japanese car.

(6) is improbable with respect to (7) (most people who own a Japanese car do not have an old Nissan); but obviously (7) is not a defeater for (6). I believe that all men are mortal; I also believe that either all men are mortal or some are not; the former is either improbable or inscrutable with respect to the latter, but the latter isn't a defeater of the former.

Given what we saw above about the way defeaters function, it is easy to see that (5) is false. Consider (6) and (7), for example. If (7) is a defeater, for me, for (6), then it would be irrational for me to continue to believe (6) after coming to believe (7). But of course this isn't so. There is nothing in the least irrational in continuing to believe (6) after realizing both that (7) is true and that (6) is improbable on it. Can we go any further? It is clear that according to (D) (7) is not a defeater of (6); but can we make a plausible conjecture as to why not? After all, sometimes, when I learn something *B* with respect to which something *A* I believed is improbable, I *do* acquire a defeater for *A*. What could make the difference? It is important to see in this case that the warrant (7) has for me is *derivative from* the warrant (6) has for me: I

learned that you had an old Nissan, and knowing that Nissans are Japanese cars, formed the belief that you own a Japanese car. But then that belief gets the warrant it has for me by virtue of being inferred (explicitly or implicitly) from my knowledge that (6); furthermore, I believe that this is the case. I therefore suggest the following. So if I believe and believe truly that the warrant a belief *B* has for me is derivative from the warrant a belief *A* has for me, then *B* is not a defeater, for me, of *A*. Here, of course, what we need is exploration and explanation of this notion of the warrant for one proposition's being derivative from the warrant for another. We have a clear case where I infer *B* from *A*, where *B* has warrant for me by way of being inferred from something that has warrant for me (as in the above case); but is that the only kind of case of this phenomenon? Further: do we really need the clause according to which my belief that the warrant *B* has for me is derivative from the warrant *A* has for me? I think not: if this belief is rational, then even if it is false, *B* will not be a defeater of *A* for me. So perhaps we could put it as follows:

> (First Principle of Defeat (FPD)) If *S* rationally believes that the warrant a belief *B* has for him is derivative from the warrant a belief *A* has for him, then *B* is not a defeater, for him, of *A*.

Aren't there still stronger principles lurking in the bushes? Suppose I *don't* infer *B* from *A*; suppose it has independent warrant for me (perhaps I know you also own a new Toyota and infer (7) from *that*). Then too, one thinks, (7) would not be a defeater for (6); but its being a defeater for (6) is not precluded by (FDP); therefore we should look for a slightly stronger principle. Given space constraints, I shall have to leave the development of that principle as an exercise for the reader.[43]

So (5) is false. But perhaps the objector can regroup. "All right, (5) is false. But my claim is really that your argument presupposes that it is true. You insisted that N&E is a defeater, for one who accepts it, for R; but the only reason you gave for thinking so is just that

---

[43]Further: suppose *S* believes *irrationally* that the warrant *B* has for him is derivative from the warrant *A* has for him: wouldn't it still follow that *B* is not a defeater, for him, of *A*?

the latter is improbable or inscrutable on the former. If you don't accept (5), what *is* your reason for thinking N&E is a defeater for R? This is an eminently fair question--although the objector seems to be overlooking the possibility that a proposition *A*'s being inscrutable or improbable with respect to *B* might be sufficient for the latter's being a defeater for the former for *some* pairs of propositions, even if not for *all* such pairs. I do want to answer this question, however, but with your permission, I'd like to defer the answer to pp. 0000 below.

B. Austere theism a defeater for R?

As you recall, the objectors claimed that austere theism--the view that we have been created by a being who is very powerful and knowledgeable[44]--raises three problems for my argument.

1. First, theism is improbable or inscrutable on austere theism.[45] Either it is improbable, on this proposition, that there is a being who is all-powerful, all-knowing, wholly good, and has created human beings in his image, or else (more likely) that probability is inscrutable. But the theist is of course committed to austere theism and (unless he has exceptionally limited powers of inference) believes it; so if he recognizes that theism *is* improbable or inscrutable on austere theism, then he has a defeater for theism.

But we already have the materials for a response to this claim. The theist, naturally enough, believes that theism has warrant for her, and she will have her candidates as to the source of that warrant. Further (she thinks), it is theism, not just austere theism, which receives warrant from these sources. Belief in God, she says, has warrant for her by way of something

---

[44]Austere theism, therefore, doesn't entail (or preclude) the proposition that we have been created in the image of an omnipotent, omniscient, wholly good person.

[45]Here for purposes of argument I ignore a complication. Many traditional theists have held that the being who is omnipotent, omniscient, wholly good, etc., is a necessary being (one that exists in every possible world) who has those properties essentially (i.e., has them in every world in which it exists): if they are correct, then theism is a necessary truth whose objective probability conditional on any proposition is 1. Furthermore if theism is taken to entail the proposition that there is a necessary being who has those properties essentially, then theism is either necessarily true or necessarily false, so that its objective probability with respect to any proposition is either 1 or 0.

like Calvin's *sensus divinitatis* or Aquinas' natural but confused knowledge of God, on the one hand, and the authority of Scripture or Church together with the internal testimony of the Holy Spirit on the other.[46] But these sources of warrant are sources of warrant for *theism*; the warrant *austere* theism has for her is derivative by way of inference from the warrant theism has for her. By the First Principle of Defeat, therefore, it won't be the case that austere theism is a defeater for theism *simpliciter*. Perhaps you will object that she is wrong in thinking theism *does* have warrant for her; but strictly speaking, that is irrelevant. What (FPD) requires is only that she be *rational*, not subject to cognitive dysfunction, in thinking that theism has warrant for her and that the warrant of austere theism is derivative from it; but surely she *could* hold that belief rationally. At any rate, she could if theism is true; but if the objector's objection depends upon the falsehood of theism then it isn't interesting in the present context.

2. The second alleged problem went as follows: first, if N&E is self-defeating in the way I claim, so is austere theism. For R is just as improbable or inscrutable on austere theism as it is on N&E. But then, by the very arguments I offered for the naturalist's having an undefeated defeater for R, one who accepts austere theism also has an undefeated defeater for R. And then (again by the same argument as with N&E) he has a defeater for any proposition he believes, including austere theism itself. So austere theism can't be rationally accepted. But theism obviously entails austere theism; and if it is irrational to accept austere theism, then it will be equally irrational to accept any proposition that entails it. Hence if my argument shows that N&E can't be rationally accepted, it follows that the same holds for theism. If the argument shows that the naturalist is in epistemic trouble, it pays the same compliment to theism; so far as this argument goes, theism and naturalism are in the same leaky boat.

By way of reply: one problem with the objection is that it relies upon the principle that

(8) If it is irrational to believe *B*, and *A* entails *B*, then it is irrational to believe *A*.

---

[46]For an explanation of how this might work, see my forthcoming (I hope) *Warranted Christian Belief*, chap. 0000.

Like (5), this principle, initially plausible as it no doubt is, needs more work. Each of nominalism and realism, for example, is rationally acceptable in certain circumstances *C*; at least one of them, however, is necessarily false; hence at least one entails just any contradiction; but not just any contradiction is such that it is rational, in those circumstances *C*, to believe it. And if you are dubious about the idea that a necessarily false proposition entails just any proposition, then note that I might be rational in believing the axioms of some formulation of arithmetic or set theory, but irrational in believing some consequence *c* of that formulation; perhaps *c* doesn't seem self-evident, I can't find any proof of it, no one I trust has so much as suggested that it is true, and I believe it only because I like the looks of the sentence expressing it. Indeed any proposition I rationally believe will have consequences I can't rationally believe.

"But surely any circumstances in which it is rational to accept theism are also circumstances in which it is rational to accept austere theism: it certainly isn't hard to follow the argument from theism to austere theism." Right; so perhaps some revision of (8) is true. Perhaps, if *B* is rationally unacceptable for me, and *A* *obviously* entails *B,* and I see that it does, then *A* is also unacceptable for me. Or perhaps not. In any event, however, the most interesting problem with the objection lies in a different direction. For what the objection shows is only that austere theism held *apart from* theism (or something similar) is irrational; it doesn't follow that austere theism accepted as a consequence of theism is irrational. Perhaps it would be irrational to hold austere theism and nothing stronger; it doesn't follow that it is irrational to accept it in the presence of theism. Perhaps theism has warrant for me in the ways mentioned above, and perhaps the only source of warrant, for me, for austere theism is by way of my inferring it from the former. Then it would be irrational for me to accept austere theism in circumstances in which I don't accept theism, just as the objector shows; but it wouldn't follow that austere theism isn't acceptable in *any* circumstances, and it wouldn't follow that my theistic belief is irrational. Any accurate revision of (8) would involve a reference to a set of circumstances: and it would imply that any set of circumstances in which it is rational to accept theism is a set of circumstances in which it is also rational to accept austere theism.

What the objector's argument shows, however, is only that it would be irrational to accept austere theism in the *absence* of theism *simpliciter.* The fact that *that* would be irrational doesn't so much as slyly suggest that a noetic structure containing theism *simpliciter* (and hence also austere theism) is irrational.

3. Third, the objector claimed that if N&E furnishes one who accepts it with an undefeated defeater for R, the same goes for austere theism: one who accepts it has an undefeated defeater for R and hence for anything else she believes. But then she has an undefeated defeater for theism--by the very same argument according to which the partisan of N&E has a defeater for N.

But there is a reply. Again, I know that you own an old Nissan; acting on the principle that it is always good to accept an additional true belief or two, I infer both that you own an old car and also that you own a Japanese car. But then I note in alarm that the second of these is unlikely with respect to the first: most people who own an old car do not own a Japanese car. In considerable puzzlement I therefore conclude that the first is a defeater for the second, and fall into a funk fretting about what I should believe here. But surely I have gone wrong; defeaters don't work that way. And again the reason is clear: both of these beliefs get their warrant from the same belief. I know that you own an old Nissan: I infer that you own an old car and also that you own a Japanese car; so the warrant each of these has for me is derivative from the warrant enjoyed by my belief that you own an old Nissan. But then neither will be a defeater of the other. More generally, I propose:

> (Second Principle of Defeat) (SPD) If *S* rationally believes that the warrant, for him, of a belief *B* is derivative from that of a belief *A*, then *B* won't be a defeater, for him, for any belief *C* unless he rationally believes that *A* is a defeater for *C*.[47]

---

[47]I owe this principle as well as FPD to correspondence with Stephen Wykstra; the precise relationship between FDP and SPD is not easy to assess. (Perhaps (as with (FPD)) the result strengthening the principle by deleting the occurrences of 'rational' also expresses a truth.)

But application to the case at hand is clear: the theist believes, perfectly rationally, that the warrant for austere theism, for him, is derivative from the warrant, for him, of theism *simpliciter*; the latter is not a defeater for R; neither, therefore, is the former.

C. Why Can't the Naturalist Just Add a Little Something?

But this leads immediately to the next objection. Austere theism, taken neat, is irrational according to my argument; the theist escapes irrationality, therefore, only because he believes something *in addition to* austere theism; but then, as Carl Ginet says, "how is it that the theist is allowed to build into her metaphysical hypothesis something that entails R or a high probability of R but the naturalist isn't? Why isn't it just as reasonable for the naturalist to take it as one of the tenets of naturalism that our cognitive systems are on the whole reliable (especially since it seems to be in our nature to have it as a basic belief)?"[48]

This sounds like an eminently fair question: is there a decent reply? Actually, however, there is more than one question here. One question is: what makes it right, appropriate, acceptable for the theist to reject the claim that austere theism is a defeater for R? That is the question we have just answered: it is because the theist rationally believes that the warrant for austere theism, for her, is derivative from the warrant theism *simpliciter* has for her, and under those conditions the former is a defeater for a belief *A*--R, for example--only if the latter is. But it isn't. There is a second question, however. What preserves R from defeat, for the theist, Ginet suggests, is the fact that the theist accepts, not just austere theism, but theism *simpliciter*. Theism can be described as T- (austere theism) plus a little something extra: T-$^+$, as we might put it. Well, why can't the naturalist do the same thing? Why can't he move to N$^+$ by adding a little something, perhaps the proposition that "our cognitive systems are on the whole reliable" (Ginet) or "we have won the evolutionary lottery"[49] or a general proposition "to the effect that the initial conditions of the development of organic life and the sum total of evolutionary processes (including ones as yet unknown or only dimly understood) were and are such as to render

---

[48]"000000", *Philosophy and Phenomenological Research*, 000 000, p. 0000.
[49]Perry, in discussion.

P(R/N&E&C&O) rather high?"[50] R(R/N$^+$&E), naturally enough, is not low or inscrutable. Doesn't this look like a good way to stave off defeat? And if this sort of thing is fair for the theist, isn't it equally fair for the naturalist?

Of course the first question is whether the naturalist who moves to N$^+$ is indeed "doing the same thing" as the theist. And the answer is that he is not. N&E, N$^+$ and R are not related, for the naturalist, the way theism, R and austere theism are related for the theist. The point was that the warrant austere theism has for the theist is derivative from the warrant theism has for her; but it is not the case that the warrant N&E has for the naturalist is derivative from the warrant N$^+$ has for him. So it is also not the case that the naturalist can properly respond by way of this *tu quoque*.

Well, perhaps: but why *can't* he move to N$^+$, thus staving off defeat of R? Better, perhaps he has believed N$^+$ all along, in which case there is no question of *moving* to N$^+$. He points out that part of his position is and has been that we have won the lottery. Granted: it is unlikely, given our evolutionary origins, that our cognitive faculties would be reliable; but of course the unlikely often happens, and, fortunately for us, it happened here. He concedes that the probability of R on N&E is low or inscrutable; but he adds that he also accepts N$^+$, on which of course the probability of R is 1. Thus, he says, he has a defeater-defeater for the defeater provided by N&E and the perception of the relation between N&E and R. As Ginet says, "Why isn't it just as reasonable for the naturalist to take it as one of the tenets of naturalism that our cognitive systems are on the whole reliable . . .?"

What shall we say of this maneuver, besides that it is excessively slick? What precisely does it come to? There are really *three* maneuvers here. Ginet suggests that the naturalist take as part of naturalism the proposition that our cognitive systems are on the whole reliable. Of course this proposition just is R, the proposition for which I claimedN&E was a defeater. Ginet's suggestion, then, is just that the naturalist could take R *to be a part of naturalism*. (He can then

---

[50]O'Connor, in "An Evolutionary Argument Against Naturalism?", forthcoming in *Canadian Journal of Philosophy*.

point out triumphantly that R is not at all unlikely with respect to the conjunction of E with naturalism so understood (N^+, we might say); N^+&E, of course, *entails* R.) That's the first maneuver; the second is Perry's suggestion that upon seeing the bearing of N&E on R, the naturalist should just add to his noetic structure the proposition L that we have won the lottery, pointing out that R is not improbable or inscrutable with respect to N&E&L.   And the third maneuver, O'Connor's, is really to add to one's noetic structure the proposition that there is a true proposition P such that R is probable with respect to N&E&P.

The first thing to see is that these procedures can't be right *in general*; if they were, every defeater could be automatically defeated. Suppose I believe the Bible is a special revelation from God and is therefore infallible: everything it affirms is true. Sadly enough, however, I read Mark Twain and H. L. Mencken in an insufficiently critical frame of mind and come to believe U, the proposition that the Bible is unreliable and full of egregious errors. (I form the opinion that a proposition's being affirmed in the Bible confers no more probability upon it than its being affirmed in any other ancient text.) U, I should think, is a defeater for any proposition I accept just on the basis of Biblical teaching. But now consider some belief *B* I do hold just on that basis--perhaps the proposition that Jesus Christ is the incarnate Son of God. U looks like a defeater for *B*. Could I defend *B* from defeat just by adding a little something to U? Ginet suggests that the naturalist add R itself to naturalism; could I analogously add *B* to U, thus moving to U*, pointing out that the probability of *B* on U* (i.e. U&*B*) is neither low nor inscrutable? Or better, since this *adding* something isn't really relevant, could I point out that I believe not merely U but *B*&U, adding that this conjunction entails *B*, and claiming triumphantly that I now no longer have a defeater for *B*? As Quine says in another connection, that is not the method of true philosophy.

Perry's suggestion is slightly different; what he thinks the naturalist should add is L, the proposition that we have won the lottery. The implied scenario is this: the naturalist comes to see that N&E is a defeater for R, but then responds by adding L to his noetic structure, thus acquiring a defeater-defeater, a defeater for his defeater of R. R is unlikely with respect to N&E,

he concedes, but now he also believes L and N&E&L; and with respect to that, again, R is neither improbable nor inscrutable. But this can't be right either. Consider the probabilistic argument from evil against theism and consider the analogue of Perry's response: "Well, I concede that the existence of God is unlikely given all the suffering the world displays, but I have a defeater for this defeater. I believe that we have won the divinity lobby, and, despite its improbability, that there is indeed such a person as God." Again, not the method of true philosophy. If you discover that you have a defeater D for one of your beliefs B, you can't in general deliver B from defeat just by noting that you believe the conjunction of D with some other proposition D^* such that $D\&D^*$ entails B. In particular, if you believe N&E and that seems initially to be a defeater of R, you can't preserve the latter from defeat just by noting that you also believe N&E&L, with respect to which the probability of R is 1.

O'Connor's suggestion is a bit different again; it is that the naturalist should add

(9) There is some true proposition P such that P(R/N&E&P) is high

to his noetic structure. But clearly this is no better than the two preceding suggestions. It would be like conceding that the existence of evil is a defeater for theistic belief, but suggesting that this defeater can be defeated by adding that you think there is some other true proposition P (theism itself perhaps?) such that the probability of theism with respect to P together with that evil is high. Once more, not the method of true philosophy.

We can see, therefore, that these responses are unacceptable. Can we see a bit deeper? Can we see *why* they are unacceptable, given the above account of defeaters and defeat? I think we can. Begin with Ginet's suggestion, that the naturalist (1) point out that he accepts the proposition that our cognitive faculties are reliable, and (2) declare that naturalism, as he understands it, includes that proposition. I say this maneuver is bootless: why? Well, consider N, the naturalist's noetic structure before he realizes the relation between N&E and R. He then comes to see that relation, i.e., to accept a proposition P about the relation between N&E and R, and moves to a noetic structure N+P. And upon coming to see P, so I claim, he has a defeater for R. Now Ginet seems to have no objection to *that* idea; his claim, instead, is that the naturalist

can restore rationality and avoid defeat simply by adding R to N, noting that R is not improbable or inscrutable with respect to (R&N)&E. But of course this changes nothing. If in fact N&E *is* a defeater for R, then N+P is an irrational noetic structure: some change is called for. The change Ginet suggests is simply that of conjoining R with N, i.e., believing the conjunction of R with N. (Perhaps the naturalist previously believed R and also believed N, but had not thought them together, and hence had not believed their conjunction.) But of course this won't help. The structure that results from adding that conjunction to N+P is obviously just as irrational as N+P is. You can't defeat a defeater just by believing its conjunction with the defeatee. Similar remarks apply to the suggestions of Perry and O'Connor: you can't defeat a defeater simply by believing that the defeatee is true even if unlikely with respect to your evidence and you can't defeat a defeater just by believing that there is some true proposition P such that the defeatee is probable with respect to the conjunction of P with the defeater.

Is there another principle of defeat lurking in the neighborhood? Perhaps so. The general problem with the suggestions of Ginet, Perry, O'Connor and the analogous suggestions I put in the mouth of that addled theist, is something like this. One acquires a defeater for a certain proposition--R, say--and then proposes as a defeater-defeater a proposition whose warrant is derivative from that of R. Thus Ginet seems to acquiesce in the suggestion that N&E (together with the recognition of the epistemic relation between them and R) is a defeater for R; but he proposes that the devotee of N&E add N^+ the conjunction of N with R, to his noetic structure. One way to construe this (and I am not sure this is the construal Ginet intends) is as the suggestion that N^+ will function as a defeater-defeater: it will be a neutralizing defeater for the defeater of R offered by N&E (together with the recognition of the epistemic relation between them). But here that can't be right. For what is the source of the warrant N^+, this proposed defeater-defeater, for me? Well, the warrant this conjunction has for me, if any, is obviously derivative from the warrant its conjuncts have for me. But one of those conjuncts just is the defeatee itself--in which case, obviously enough, we don't have a successful defeater-defeater. Accordingly I propose

(Third Principle of Defeat) (TPD) If *D* is a defeater of *B* for *S*, then for any belief *B**

of *S*, if *S* rationally[51] believes that the warrant *B** has for her is derivative (wholly or

partly) from the warrant *B* has for her, then *B** is not a defeater-defeater, for *S*, of *D*.

If Ginet's suggestion is that N^+ will function as a defeater-defeater (a defeater for the defeater

N&E provides for R), then he is mistaken: obviously the warrant N^+ has for him is partly

derivative from the warrant R has for him. Similarly for Perry: what he proposes is that the

naturalist just add to his set of beliefs the proposition that we have won the lottery, that R is

indeed unlikely, but nonetheless true. But clearly the warrant (if any) this belief has for the

naturalist is derivative from the warrant R has for her: hence (given that the naturalist sees that it

is so derivative in that way) by TPD it can't function as a defeater-defeater. And the same goes

for O'Connor's suggestion that there is some true proposition P such that the probability of R

with respect to N&E&P is high: here two the warrant this proposition has for him is apparently

derivative from the warrant R has for him. At least he has suggested no other source for that

warrant, and it is hard to see what other source there could be.

D. R Beyond Defeat?

But perhaps this is not the way to understand O'Connor: perhaps he isn't proposing that

proposition P as a defeater-defeater at all. Here is another possibility. O'Connor grants that *if*

N&E really is a defeater for R, then the Sosa-Ginet-Perry-O'Connor maneuvers are futile. But

there's the rub: *is* N&E (together with a recognition of their epistemic bearing on R) a defeater

for R? O'Connor's suggestion (on this reading) is that it is not. His idea is that R has a great deal

of *original* or *intrinsic* warrant, for us, as we might put it. That our cognitive faculties are

generally reliable seems to be something like a natural presupposition of our entire cognitive

lives.[52] Furthermore, R has this warrant in the basic way; R's warrant does not depend upon its

---

[51]Again, can we strengthen the principle by dropping this word?

[52]Here Bas van Fraassen's "Belief and the Will" *Journal of Philosophy*, Vol LXXXI No. 5 (May
l984) is instructive. Van Fraassen proposes a principle, "(Reflection)", according to which my
subjective probability for a proposition, conditional on the supposition that at some time *t* my
probability for it will be *n*, should be that same number *n*. Perhaps there are legitimate doubts

evidential relationship with other beliefs. Recall the quotation from Ginet, who says "Why isn't it just as reasonable for the naturalist to take it as one of the tenets of naturalism that our cognitive systems are on the whole reliable (especially since it seems to be in our nature to have it as a basic belief)?" We can raise questions here about whether R is an explicit basic belief, or only implicit, what is it for a belief to be implicit and the like; suppose we forgo the questions and concede what in any event seems likely: that R is rationally accepted in the basic way, and, so taken, has much warrant for us. But if we believe R, of course, and also think our cognitive faculties have developed by way of evolution, we will conclude that that development has taken place in such a way that R is true:

> It does not seem at all objectionable [for the partisan of N&E] to reason thus: I believe R without having any ultimately non-circular reasons for doing so and know that I am nonetheless rational in so believing. Therefore, it is reasonable for me to believe (in the absence of evidence directly to the contrary) that the sum total of factors responsible for me and other human beings having the cognitive equipment that we do is such as to render R fairly probable. I take myself to have sufficient reasons for believing E very strongly, and N fairly strongly, and I note that I'm not in a position to give much of an estimate of the value of $P(R/N\&E\&C^{53})$. Therefore, I seem to be entitled at this point to suppose that other factors obtained that together with N&E&C render R probable (007).

This passage suggests the following. The naturalist first notes that R has a great deal of warrant for him, warrant it has in the basic way. This warrant is so great, furthermore, that N&E together with Q, the proposition specifying the epistemic relation between N&E and R, doesn't

---

about the *general* truth of Reflection (see my *Warrant: the Current Debate* (Oxford: 1993) pp. 000-000), but it certainly seems plausible where t = the present. The question "What is your personal probability for the proposition that p, given that your personal probability for p is n?" can only be taken as a joke or an insult.

[53]See footnote 11.

suffice to defeat R: R has so much intrinsic warrant, we might say, that it can't be defeated. But then the rational thing to think is that

> (10) There is some true proposition P reporting those "other factors" which is such that R is probable on its conjunction with N&E.

On this understanding, O'Connor is not proposing (10) as a defeater-defeater (N&E&Q don't constitute a defeater for R, so no defeater-defeater is needed); the existence of such a proposition is instead a perfectly proper deduction from R and E.

By way of reply: suppose we agree that we do ordinarily believe R in the basic way, and are furthermore perfectly rational in so doing. Let's also agree that R does has warrant and perhaps a great deal of warrant, when it is taken as basic. Still further, we can add that R plays a unique and crucial role in our noetic structures: if we are reflective and come to doubt R, we will be in serious epistemic trouble.

But it doesn't follow that I can't acquire a defeater for R: clearly I can. Suppose I assume in the ordinary way that my cognitive faculties are reliable, but then come to suspect and finally to believe that I am insane. Once I see the connection between this belief, the belief that I am insane, and R, I have a defeater for R. This example will do the job nicely: but if you have a flair for the dramatic, preferring Cartesian demons and brains in vats, we can easily construct examples to suit. If I come to believe that I am a victim of a deceptive Cartesian evil demon, then I have a defeater for R, as I would if I came to believe that I am in the clutches of Alpha Centaurian cognitive scientists who are using me as the subject for a cognitive experiment in which they induce extensive and bizarre false belief in order to see how the noetic structure reacts. The fact that R has warrant in the basic way doesn't shield it from the possibility of defeat. N&E&Q, furthermore, does indeed seem to be a defeater for R: these alleged facts about the origin of my cognitive capacities, just as in the Cartesian demon and brain in a vat scenarios, are obviously the relevant considerations.

But then what accounts for the inclination to think R has a special status, a colossal degree of warrant, so much warrant, in fact, that it can't be defeated? I think the answer is to be

found in the following neighborhood.   R, we might say, is in the human design plan for at least two different reasons.[54]   In the first place, R has warrant in the ordinary way: it is in the design plan because as a matter of fact it is true, and the design plan, being aimed at enabling us to know truth, includes it in order to enable us to know *that* truth. The design plan includes the production and sustenance of R (under the normal conditions) for the same reason that the design plan includes the production and sustenance of any other true belief. But secondly, this part of the design plan isn't aimed only at our knowing the truth about the subject matter of R (unlike the part of the design plan governing the belief that 2+1=3); it is also aimed at making it possible for us to carry on our entire noetic enterprise. (If I come to doubt R, if I think about it but do not believe it, I'll be headed for epistemic disaster.) So R's being present is certainly a matter of proper function. But the design plan here, with respect to this *second* purpose, isn't aimed *directly* (as we might put it) at the production of true beliefs. With respect to this second purpose, it is aimed *indirectly* at the truth, by being aimed at making it possible for us to carry on our whole noetic enterprise (which as a whole is aimed at truth) in a satisfactory way.

Accordingly, the design plan has a double aim or purpose with respect to R; but only the first of the two aims is relevantly connected with warrant. (Compare cases of *tradeoffs and compromise*, where the doxastic response to a given circumstance isn't aimed directly at the truth, but at some best compromise between the aim at truth and the satisfaction of other constraints having to do with mobility, brain size, etc. (see WPF pp. 38 ff.). A belief has warrant for you only if the segment of the design plan governing its production is *directly* rather than indirectly aimed at the production of true beliefs (WPF, p. 40); hence R doesn't acquire colossal, undefeatable warrant by virtue of its serving that second purpose. But it is this second function of R--the role it plays in enabling us to carry on our whole noetic enterprise--that makes R seem

---

[54]If we think of an organic design plan as a set of triples of circumstance, cognitive response, and purpose or function (see WPF, pp. 22 ff.), then the relevant triples involving R would mention at least *two* purposes or functions.

so essential to our cognitive lives. And it is indeed essential; but that isn't sufficient for its being undefeatable.

E. The Dreaded Loop

The objector complained that there were two problems with what I originally said on this head. First, in WPF (p. 235) I unwisely followed Hume and Sextus Empiricus in arguing that the devotee of N&E, if rational, will fall into the following sort of diachronic loop: first, he believes N&E and sees that this gives him a defeater for R; so he stops believing N&E; but then he *loses* his defeater for N&E and R; so presumably those beliefs then come flooding back. But then once again he has a defeater for them, and withholds them; and so on, round and round the loop. So what you really have is a loop where N&E keep getting alternately defeated and reprieved-- i.e., at $t1$ it is defeated, at $t2$ undefeated, at $t3$ defeated, and so on. And I went on to say that this situation gives him an ultimately undefeated defeater for R. But (and here comes the objection) even if he got into such a diachronic loop, it wouldn't be the case that he would have an ultimately undefeated defeater for R; what he would have instead is a defeater that is not ultimately defeated. And second, why think rationality requires that he get into this loop in the first place? Can't he see in advance what's going to happen?

The objector is right on both counts. Let me penitently offer the following correction. The devotee of N&E, if he is canny (or reads what I am about to say) sees that there is a certain *synchronic* structure here. He sees that there is an infinite series of potential defeaters and defeatees; but he needn't himself doggedly plod around any diachronic loop. To see the relevant structure, consider a similar but simpler case: imagine a skeptic who comes to believe -R, that his faculties are not reliable. What he should see (what in any event *we* can see) is that there is then an infinite series of propositions related in an interesting way. At the first level, there is -R, which he believes. But there is a connection between -R and any other belief he has, including -R itself: if -R is true, then any belief he has, including -R itself, will be unreliably formed. But *that* belief--the belief that -R is unreliably formed-- gives him a defeater for -R. Suppose we let '-R(p)' express the proposition that p is unreliably formed. Then at the second level we have -R

and -R(-R), which is a defeater of -R. But then at the third level we have -R and -R(-R(-R)), which is a potential defeater of -R(-R), the defeater of -R; and so on. Perhaps we can schematically represent the structure as follows:

level 0 -R

level 1 -R and -R(0) (i.e., -R(-R)) (which is a potential defeater for level 0, i.e., -R)

level 2 -R and -R(1) (which is a potential defeater for level 1 and a potential defeater-defeater for level 0)

level 3 -R and -R(2) which is a potential defeater for level 2 and a potential (defeater-defeater)-defeater for level 0

.

.

.

level n -R and -R(n-1) (which is a potential defeater for level n-1)

.

.

.

There is an infinite series of propositions here, but of course not an infinite series of defeaters and defeatees. *D* is a defeater of *B* only if *D* is a defeater of *B for someone*; *D* is a defeater of *B* for someone *S* only if *S* believes both *D* and *B*; but no one could believe all the propositions in this infinite series. Still, we might say that each member is a *potential* defeater for the preceding member, in that if someone believed n-1, and saw -R(n-1), he would have a defeater for n-1.

And the important point here is this: in thus acquiring a defeater for level 1, i.e., in acquiring a defeater-defeater for the defeater of level 0, one does not lose one's defeater for level 0. As long as the skeptic believes -R, she has a defeater for -R: -R(-R). And she has this defeater even if she also has a defeater for -R(-R).   This is, of course, extraordinary: ordinarily, if one acquires a defeater-defeater for a belief *B*, i.e., a defeater for a defeater of *B*, one no longer has that defeater for *B*--or else its defeating power is neutralized. But not so here. The

difference is that here the original defeatee shows up at every subsequent level. When that happens--when, roughly speaking, every defeater in the series is really the defeatee plus a bit, the defeater-defeater doesn't nullify the defeater. The defeater gets defeated, all right, but the defeatee remains defeated too. Accordingly, any time at which the skeptic believes -R, he has a defeater for -R—even if he also has, at that time, a defeater for that defeater. Skepticism of this kind, then, really is self-defeating, even if it is also the case that the skeptic has a defeater for his defeater.

But the same goes for N&E. At level 0, we have N&E&Q (Q = the proposition stating the epistemic relation between N&E and R). N&E&Q, however (for any *S* who believes it), provides *S* with a defeater for R and hence for any other proposition *S* believes. Let 'N&E&Q(p)' denote the belief that *p* is formed by faculties such that the probability of their reliability on N&E is either low or inscrutable. Then perhaps we can represent the relevant structure as follows:

level 0 N&E&Q,

level 1 N&E&Q and N&E&Q(N&E), which is a potential defeater for N&E&Q

level 2 N&E&Q and N&E&Q(1), which is a potential defeater for level 1

.

.

level n N&E&Q and N&E&Q(n-1), which is a potential defeater for level n-1

.

.

.

Here at each level there is the proposition N&E&Q; this proposition, if believed, is a defeater for itself, and each level of the structure, if believed, is a defeater for the preceding level. The central point here, just as with the simpler skeptical structure, is that the devotee of N&E has a defeater for N&E at any time at which he believes that proposition, and sees that R is improbable or inscrutable with respect to N&E, so that if he believes N&E he has a reason for withholding any of the beliefs he accepts, including N&E itself. He has such a defeater for N&E even if at

that time he also has a defeater for that defeater. His problem, after all, is that N&E gives him a defeater for *everything* he believes.

By way of conclusion then: the evolutionary argument against naturalism is subject to several intriguing objections.   Evaluating these objections requires taking a closer look at the nature of defeaters and defeat. What that closer look reveals, I think, is that the evolutionary argument emerges unscathed. Naturalism alone may (or may not) be tenable, and the same goes for the view that we have evolved by way of the mechanisms to which contemporary evolutionary theory directs our attention; the conjunction of these two propositions, however, can't rationally be accepted.[55]

--Alvin Plantinga

December, 1994