On Plantinga on Belief in Naturalism[1]

Troy Cross

Introduction

Naturalism, according to Alvin Plantinga, is the thesis "that there is no such person as God or anything at all like God" (2011b, 435; see also, 2002a, 1; 2011a, *vix*), and one could be forgiven for thinking that what Plantinga calls his "Evolutionary Argument Against Naturalism", or "EAAN", which he has refined and defended for over two decades, is an argument against that thesis (1991, 1993, 2002a, 2002b, 2011a, 2011b). But that would be a mistake. The conclusion of the EAAN is not about the *truth* of naturalism but about the *rationality of believing* naturalism to be true: it is that one cannot rationally and reflectively believe that contemporary evolutionary theory is correct about our origins while also believing that neither God nor anything like God exists.[2]

The difference between truth and rational belief is important here. Consider Moore's paradoxical sentence: "I don't believe it's raining but as a matter of fact it is" (Moore, 209). Nothing prevents that sentence, or others like it, from being true. If you

[1] For instructive conversations on this topic I am indebted to John Bang, George Bealer, Mark Bedau, Eliyah Cohen, Bryan Cross, Augie Faller, Emma Handte, Elad Gilo, Nick Gigliotti, John Hare, Jordan Horowitz, Paul Hovda, Alexander Pruss, Margaret Scharle, Derek Schiller, Mackenzie Sullivan, and Carol Voeller. I also wish to thank audiences at Reed College and the University of Nebraska Omaha where I presented early versions of some of this material. Anthony Nguyen gave me many helpful comments on a draft. Finally, two philosophers' work on the EAAN deserves special thanks: Ernie Sosa, whose articles I never cite explicitly in the paper, but which shaped the general way I think about the EAAN, and Omar Mirza, whose conjecture that we will advance in our understanding of the argument only by a closer examination of the "XX pill" case (2011, 86) prompted my dilemma for the argument by way of that case.

[2] Some early presentations paired the EAAN with a "preliminary" argument that naturalism is false though that argument has disappeared in later iterations (Plantinga, 1993, ch. 12). Perhaps it was this preliminary argument that prompted the name.

have no belief about whether it's raining, then it's either raining but you don't believe it to be raining or it's not raining but you don't believe that it's not raining. Either way, some Moore sentence is true. Nevertheless, it would be irrational to believe one of them. And as it goes for Moore sentences, so it goes for naturalism and evolution according to Plantinga. The EAAN purports to show *not* that naturalism and evolution are false or improbable, but that *even if they are tru*e, and in fact, *even if they are overwhelmingly supported by our evidence*, their conjunction cannot rationally be believed.[3]

This is surprising to say the least. The view that Plantinga deems inherently irrational, *viz*., the conjunction of naturalism and evolution, does not immediately strike us as paradoxical in the way that Moore sentences do. To the contrary, it is endorsed by many working scientists and philosophers.[4] Nor is the conjunction of naturalism and

---

[3] The mechanism employed by the EAAN, to be explained shortly, is very different from the mechanism underlying the rational incoherence of Moore sentences. Naturalism and evolution are not directly *about* belief, so there is no *logical* connection between the state of *believing* the conjunction of naturalism and evolution, and the *content* of that belief, and therefore, no Moore-style paradox. Likewise for Fitch-paradox sentences of the form "*P*, but it will never be known that *P"* (Fitch, 1963). Supposing one should not believe what one does not know, all such sentences, even if true, cannot be rationally believed. Or relatedly, consider a generalized Moore sentence, or a weakened Fitch sentence: "*P*, but it will never be believed by anyone that *P*". All of these paradoxes involve sentences whose contents *involve belief*, and that is key to understanding their paradoxicality. They all display a kind of Heisenberg-uncertainty-like phenomenon, where, just as measuring a quantum state disrupts the state being measured, so, believing the truth in question in some way disrupts that very truth. This feature also distinguishes these paradoxes from Plantinga's argument, because naturalism and evolution do not directly involve belief or non-belief in their contents and do not turn on the interference of measurement, so to speak. But what the paradoxes do have in common with the EAAN is the *non-alethic* basis for their paradoxicality. On this score, note further that Moore and Fitch sentences are equally irrational for highly reliable and highly knowledgeable individuals (short of omniscient ones, for whom there would be no true sentences of that form) as for less reliable and largely ignorant individuals. Their paradoxicality, in other words, has nothing to do with their probability. Likewise for naturalism. Its putative paradoxicality has nothing to do with its probability, or with the evidence for or against it.

[4] Pinning down a precise estimate of the popularity of naturalism and evolution is difficult. According to a Pew Research Center survey, 98% of members of the American Association for

evolution simply a widely held *personal* belief, something like a cultural identity or stylistic preference, of people who happen also to be scientists. Evolutionary theory is foundational, indispensable, for biology and related sciences, and across the sciences it is accorded respect and adherence.[5] Naturalism, likewise, lays claim to a special *normative* status within scientific institutions and practice. Plantinga writes:

> Naturalists pledge allegiance to science; they nail their banner to the mast of science; they wrap themselves in the mantle of science like a politician in the flag. They confidently claim that naturalism is part of the "scientific worldview," and that the advent of modern science has exposed supernaturalism as a tissue of superstition – perhaps acceptable and perhaps even sensible in a prescientific age, but now superseded" (2011a, 307)

Perhaps Plantinga has certain zealots of naturalism (Richard Dawkins, Daniel Dennett, and Sam Harris) in mind here. Not all naturalists, surely, are pledging allegiance, nailing

---

the Advancement of Science believe that "humans have evolved over time", and 41% believe neither in God nor a higher power or universal spirit (Pew 2014; Pew, 2009). Atheism is also prevalent among professional philosophers, 73% of whom "lean towards" atheism according to a survey by *Phil Papers* (Bourget and Chalmers). Presumably, also, many of these atheistic scientists and philosophers would affirm the theory of evolution. (Evolution-rejecting atheists are rare. Thomas Nagel comes to mind, though his argument for the inadequacy of evolutionary theory has found few sympathizers (Nagel, 2012).) The more serious worry for quantifying the popularity of Plantinga's target belief is that the description, "anything like God" in the characterization of naturalism may simply be too open-ended to allow for a determinate answer. As we will see, Plantinga means to include as too God-like, the existence of anything not material or "determined by" the material, so dualists, perhaps Platonists, semantic primitivists, normative primitivists, *may* all count as anti-naturalists. Still, despite the indeterminacy of its meaning, or even given one its stronger, more exclusivist readings, I think a great many scientists and philosophers would probably agree that "nothing like God" exists.

[5] Evolution is not, in any case, Plantinga's target. He himself endorses contemporary evolutionary theory, or at least claims to do so (2011a, 310). If one builds into evolutionary theory the assumption of *true randomness* in the introduction of variation in organisms, then it is not clear he does, in fact, endorse the theory of evolution, at least on some ways of understanding what true randomness entails.

banners, wrapping themselves in flags, and so forth. But he is right that certain prominent naturalists do so, and without much resistance from other naturalists. In any case, the EAAN thus threatens to locate in humanity's epistemic blindspot not only a widely held belief, but a doctrine that presumes a serious and weighty kind of epistemic authority.

Of course, such a provocative argument has drawn the critical attention of dozens of philosophers over the years, some pointing to what they call "serious errors" (Fitelson and Sober, 1998, 115) and others claiming the argument outright "fails" (Talbott, 164) or at least gives the naturalist "nothing to fear" (O'Connor, 134). But Plantinga continues to deftly defend and conveniently revise the EAAN in response to his critics, with the result that, decades after its introduction, it remains, as Stephen Law notes, "... one of the most widely discussed arguments targeting philosophical naturalism" (2012), and as Omar Mirza says: "… it does not seem as though we are nearing a resolution of the debates that Plantinga's argument has provoked" (2008, 126).

My primary goal in the present paper is to set out a new critical response to the EAAN. I shall argue that the EAAN turns on a key analogy -- the argument simply will not work without it -- but that the analogy harbors an ambiguity. However we choose to resolve this ambiguity into something more precise, a hopeful strategy for the naturalist will emerge.

I have secondary goals in the paper as well. I want to set out the best and most contemporary version of the argument, distinguishing the parts that have survived from those that have not; I want to strengthen the argument where I can; I want to review, and

where possible improve, some earlier objections (certainly not all), folding them into my own; I want to argue that if a naturalist is convinced by Plantinga's argument, there is no rational pathway for her to arrive at a belief in anti-naturalism, not even a pathway guided by prudential rationality;[6] and finally, I want to explain why debate over the EAAN has not, and will not, resolve itself anytime soon even given my "hopeful" strategy.

To preview this last point briefly, my explanation for the longevity of the EAAN is that its premises turn on many underlying philosophical issues including but not limited to the following four:

> (i) whether we have already discovered, or will discover in the near future, a compelling, independently motivated naturalized theory of content that roughly matches our pre-theoretic beliefs about belief;
>
> (ii) how our best philosophical theories, throughout their stages of development, should constrain our rational beliefs about the subject matters of those theories;
>
> (iii) how our beliefs should be shaped by higher-order evidence, and in particular, evidence about the *origins* of our beliefs; and
>
> (iv) whether "internal" states are in any way privileged in fending off a certain kind of skeptical attack.

These four philosophical questions *do* have answers that are individually somewhat plausible, and which, if true, would make it at least possible that Plantinga's argument is sound. But there are other answers, and sets of answers, which are to my mind much

---

[6] This sets Plantinga's argument apart even from Pascal's Wager, where at least practical reasoning guides the non-believer along the road to eventually acquiring theistic belief. The EAAN is, for this reason, the most Calvinist "apologetic" argument yet conceived: it silences the naturalist, but cannot offer her a rational hope for any fate other than skepticism.

more plausible, though still controversial, and which if true would definitely render Plantinga's argument *un*sound. Thus, even if some readers are convinced by my proposed reply to the EAAN, the argument will be no sooner be either debunked or defended to *everyone's* satisfaction than the broader philosophical issues (i)-(iv) on which it hinges are resolved in a way that garners the same degree of universal acclaim.

The EAAN

I begin with a sketch of the argument. Following Plantinga, let "N" abbreviate *naturalism*, the proposition that neither God nor anything like God exists (2011a, 317). Let "E" abbreviate *evolution*, the proposition that "we and our cognitive faculties have come to be in the way proposed by the contemporary scientific theory of evolution" (2011a, 317). Let "R" abbreviate *reliability*, the proposition that our cognitive faculties are generally reliable (2011a, 317).

The EAAN, in outline, is as follows (2011a, 344-345):

1. P(R|N&E) is low.
2. Anyone who accepts both N&E and that P(R|N&E) is low has a defeater for R.
3. Anyone who has a defeater for R has a defeater for any other belief she thinks she has, including N&E itself.
4. If one who accepts N&E thereby acquires a defeater for N&E, N&E is self-defeating.
5. Conclusion: N&E can't rationally be accepted.

The first premise is known as the "probability premise". It says that the probability of our belief-producing faculties – memory, perception, intuition, inference -- being generally reliable, given naturalism and evolution, is low. The argument for the probability premise begins with the assumption that naturalism entails *materialism*, the thesis that human beings are material objects (2011a, 318-320). Plantinga, developing a worry first expressed by Darwin himself, argues that there is no reason why, given that human beings are mere material objects and beliefs are mere states of the nervous system, random processes introducing variation in organisms would differentially *bring about* true-belief-producing faculties over false-belief-producing faculties in the first place, or why natural selection would *select against* false-belief-producing ones after such faculties have been brought about (2011a, 316). There is simply no reason to think, given our material nature and the mechanisms that brought us about, that we have reliable belief-forming faculties.

In evaluating the probability premise, Plantinga asks that we imagine creatures "like ourselves" in that they have beliefs and make inferences, but about which we know only that they resulted from natural selection in a world absent anything like God (2011a, 325, 327). We are supposed to surmise that it is highly improbable that *they* are, in general, truth-believing. Following that probability judgment, we should, according to Plantinga, set the probability that *we ourselves* are generally reliable equal to the probability of reliability we assign to these hypothetical creatures. Thus, given only naturalism and evolution, Plantinga thinks the reliability of our own belief-producing faculties is improbable.

The second premise, known as the "defeat premise", says that acknowledging the improbability of our reliability given naturalism and evolution while at the very same time believing naturalism and evolution "defeats" the belief that our faculties are reliable. In "defeat"-free language, the premise says that seeing the truth of the probability premise and being a naturalist and evolutionist, one cannot continue rationally to believe that our faculties, including one's own faculties, are generally reliable.

Plantinga argues for the defeat premise by means of an analogy: if you believed yourself to have taken a pill, which he calls "XX", which you believe makes the vast majority of its ingesters globally unreliable, then you could not rationally believe your own belief-producing faculties to be reliable (2011a, 342). In just the same way, the naturalist cannot believe her own faculties to be reliable, given that she acknowledges the truth of the probability premise. Doing so would be like believing oneself to have won the evolutionary lottery.

Finally, Plantinga argues that once we have a defeater for the belief that our belief-producing faculties are generally reliable, we have a *universal* defeater, a defeater for all of our beliefs, including naturalism and evolution, i.e., if we do not think we our belief-producing faculties are truth-conducive, we cannot rationally believe any of the fruits of those (probably) unreliable faculties, which means we cannot rationally believe anything at all. He concludes, thereby, that evolution and naturalism are "self-defeating", that believing them actually makes them, along with everything else, irrational to believe.

The Probability Premise: P(R|N&E) is low.

The argument for the probability premise has shifted over the years, but in ways that are not always easy to track; Plantinga does not explicitly disavow his earlier versions, but merely updates with new ones. I will mention in passing some of the earlier considerations but focus my attention on the latest editions as best I can (2011a, 2011b).

All of the arguments, old and new, work in the same *general* way. They all begin by assuming materialism is a consequence or component of naturalism. They all suggest that we ask not about ourselves first of all, but about hypothetical believing creatures who evolved, like us, in a godless, materialistic universe. Then, they all argue that we should think it improbable that *those* creatures are reliable, because all we know about these creatures is that they evolved through the introduction of random variation and natural selection, and we should see that knowing only that those beliefs are adaptive tells us nothing about whether they are also true. A believing organism may have all false beliefs or all true beliefs or any ratio of true-to-false beliefs whatsoever, and be equally well adapted. The truth or falsehood of a given belief, or the prevalence of truths in a noetic system, says Plantinga, is totally unrelated to the fitness of the organism. Finally, we should think it no more likely that *we ourselves* are reliable than that these *hypothetical believing creatures* are reliable.

Think of natural selection as a kind of giant filter. Random combinations of organic material arise spontaneously, some of them reproducing other organisms like themselves. Chance events like storms, meteors, earthquakes, and so on, will, for no good reason, kill off some populations of reproducers and others will remain, "selected" purely by chance. Over a vast range of space and time, however, the organisms that

survive and reproduce at a differentially higher rate than competitors in their ecological niche*, as a matter of how they are constructed*, will become more prevalent.

Suppose one is told to imagine a hypothetical organism that results from such a process, but told nothing more, and asked what is likely true *of it*. Given that better reproducers are more prevalent than poor reproducers in environments shaped by evolution, choosing an organism at random, one can assume there is a high probability of having chosen a better reproducer. One can assume that if an organism was *extremely* poorly suited to its environment -- incapable of reproducing at all, or reproducing at an extremely low rate relative to similar organisms -- it would be rare, and therefore, having chosen at random, one is unlikely to be thinking of such an organism. Now, given the variety of possible ecological niches, it is extremely difficult to say what is or isn't maladaptive, generally.[7] And given that we have not specified anything whatsoever about our hypothetical creature's environment, we can determine almost nothing about what it likely is and what it likely isn't. But natural selection will filter out, for instance, creatures who opt out of reproduction altogether. They might arise spontaneously, but they will not continue on. So our hypothetical creature is not likely one of those failed experiments that is much like an organism in function, say, except for reproduction.

---

[7] I have said that a low reproductive rate is disadvantageous, for instance. But of course many species thrive with very low reproductive rates, because there are also costs to high reproductive rates. And it all depends on what competing organisms are doing. Speculating about what is and is not probable, given only the information that a thing is the result of natural selection, from the armchair, is nearly *hopeless*. From the armchair, few would guess, for instance, that the lifecycle of the cicada would provide a selective advantage. Or, choose your favorite member of a symbiotic pair. Without ecological context, how could their actual traits -- which are only adaptive within the symbiotic relation -- be assigned anything but an absurdly low *objective* probability?

Now, Plantinga's suggestion is that we add one more feature to our hypothetical creature: it has *beliefs*. That is our only additional piece of information. And now we ask whether there is any reason to think, given its origins, that its *beliefs* are more likely true than false. His argument is that we have no reason to think its beliefs are mostly true, because, on the assumption of materialism, natural selection does not have a *filter* for false beliefs. False beliefs, he thinks, are not in any way fitness diminishing, nor are true beliefs in any way fitness enhancing. *The truth or falsehood of belief, on the assumption of materialism, just makes no difference whatsoever to the fitness of the believing organism.* That is the key premise in the argument for premise 1. Given that each belief of an adapted, believing organism is no more likely true than false, the likelihood of that organisms having a high proportion of true beliefs (R's being true of it) is low.

Now, common sense tells us that an organism's *knowing* how to get nutrition, how to flee predators, how to shelter from the elements, and how to reproduce will enhance its fitness. And what is false is not known, of course. And Plantinga agrees that knowing these things is *actually* adaptive and not knowing them is maladaptive. But he contends that *on the assumption of materialism*, knowing these things is no more adaptive than being ignorant of them.

True beliefs are no more adaptive than false beliefs

Plantinga's rationale for thinking that, given materialism, the truth of a belief has nothing to do with its adaptivity has shifted over the years. He has employed three distinct, though complementary, arguments:

(i) Given materialism, *epiphenomenalism* about belief -- that beliefs do not cause behavior at all -- is probable. If beliefs do not cause behavior, then *true* beliefs certainly cannot cause differences in the reproductive success rates of those who have them, i.e., true beliefs cannot be adaptive (2002a, 6).

(ii) Even if beliefs do have causal powers, on materialism it is likely that *semantic epiphenomenalism* is true, and the *contents* of beliefs are causally inert, and it is instead the neurophysiological basis of beliefs, or the neurophysiological properties of beliefs (depending on whether one is a non-reductive or reductive materialist) that do all of the causal work. Therefore, again, true beliefs cannot, by virtue of their *contents*, and hence by virtue of having *true* contents, confer selective advantage (2002a, 6-7; 2011a, 336-339).

(iii) Even if beliefs *do* cause behaviors *in virtue of their contents*, still, *true* beliefs are no more likely to confer selective advantage than *false* ones (2002a, 7-8; 2002b, 218; 2011a, 330-334; 2011b, 441-442)

I say the arguments are complementary because how likely R is, given N&E will be a function of the Law of Total Probability. If, given N&E, epiphenomenalism is even 10% probable, but R is highly improbable given N&E and epiphenomenalism, that may still influence the probability of R given N&E somewhat, even if only a little. And the same will go for semantic epiphenomenalism. If semantic epiphenomenalism is 20% likely, but R is extremely improbable given semantic epiphenomenalism, then the probability of R may still end up lower than it would have been without considering the probability of semantic epiphenomenalism. So a naturalist might not be convinced that

epiphenomenalism or semantic epiphenomenalism are *true*, yet still, considering those possibilities might be relevant to her estimate of the probability of R, given N&E.

But the first sort of argument, the appeal to epiphenomenalism, does not appear in Plantinga's most recent work (2011a; 2011b) and I will follow him in largely ignoring it here. No materialist who thinks epiphenomenalism *highly* improbable should be concerned by it. By contrast, the third kind of argument, that false beliefs are as adaptive as true ones, appears in *all* presentations of the argument, though the sub-arguments in favor of (iii) have also varied, as I will explain shortly. The second kind of argument, (ii) the argument by way of the probability of *semantic* epiphenomenalism, does figure centrally in his recent book, *Where the Conflict Really Lies* (2011a), so I feel I cannot ignore it. But he offers, to my mind, a convincing reply to (ii), in "Content and Natural Selection" (2011b, 438), a reply which echoes a much earlier response to Jerry Fodor (Plantinga, 2002b, 218; Fodor, 2002) though it is not framed by Plantinga as a reply to himself, and though his argument for semantic epiphenomenalism is never explicitly disavowed. Additionally, (ii) might be argued in a way that Plantinga does *not* pursue in his recent work, but which avoids these difficulties. The status of (ii) is thus highly unclear. So, to recap, I am ignoring (i), which does not appear in recent versions of the argument, and before getting to the arguments for (iii), which are still very much live, I would like to consider (ii), which appears in some recent presentations and is rejected (though not explicitly) in others.

(ii) The argument for semantic epiphenomenalism given reductive materialism

In *Where the Conflict Really Lies*, Plantinga argues for the probability premise by beginning with the following claim: materialists will be either be *non-reductive* materialists who think the mental is "determined by" or "supervenes on" the physical, or else they will be *reductive* materialists who are *type identity theorists* about the mind (2011a, 322-323). He discusses the threat of semantic epiphenomenalism under the sub-possibility of *reductive* materialism, and only under that sub-possibility (2011a, 335-339).

His argument for (ii) goes as follows. According to reductive materialism, beliefs are type-identical to neurophysiological properties: beliefs are identical to massive disjunctions of conjunctions of such properties (2011a, 323-324). However, even though belief states are *identical* to neurophysiological states, they do not cause behavior *in virtue of* their contents, but *in virtue of* their neurophysiological features (2011a, 335-339).

This is a familiar sort of criticism of *token* identity theories (for an overview, see McLaughlin, 1989). Suppose, for example, that physical properties are not identical to mental properties, and that the former are law-governed while the latter are not. Suppose causal responsibility for a given event is determined by what falls under so-called "covering laws". Then even if every mental event (token) is identical to some physical event (token), it is hard to see how mental events are causal *by virtue of* their mental properties, for there are, by supposition, no psychological laws, though there are physical laws. It seems that these token events, which are both mental and physical, are doing all of their causing *by virtue of* their physical properties and through the physical laws that

govern them, while their mental properties sit idly by, somehow tacked onto the relevant event tokens, but contributing nothing to their causal transactions.

Plausible enough. But in Plantinga's argument for semantic epiphenomenalism we have not assumed *token* identity, but rather, *type* identity -- he presents it under the supposition of "reductive materialism," defined as type identity theory -- and it is simply hard to square semantic epiphenomenalism (semantic *type epiphenomenalism*) with Leibniz's Law.[8] What is Plantinga's argument, exactly? It does not go by way of a covering-law model of causal explanation. Rather, it is based on the following *counterfactual* for a given belief: "If the belief had had the same [neurophysiological] properties but different content, it would have had the same effect on behavior" (2011a, 337). Plantinga concludes from this true counterfactual that the adaptivity of belief is *really* due to its neurophysiological properties, and not its content (2011a, 337).[9] It is a kind of counterfactual test for causal relevance: if you start with two candidate causes for an effect and note that the effect would remain while changing one, but not while changing the other, you have found the genuine cause of the effect, namely, the candidate on which the effect counterfactually depends.

In "Content and Natural Selection", however, while discussing reductive materialism, Plantinga takes a *different* stance on the same counterfactual: "if [the belief] had had the same NP properties but different content, then it would have made the same

[8] Of course type identity will *entail* token identity, but in standard usage, "token identity" means *mere* token identity: token, but not type, identity.
[9] Plantinga recognizes this is not just a counterfactual, but a counterpossible, given reductive materialism, but he argues that counterpossibles are also very often informative, and he still takes it to show that the content of beliefs is doing no real causal work (2011a, 338).

causal contribution to behavior," (2011b, 438). In this article, he judges the counterfactual to be "of dubious relevance" to semantic epiphenomenalism, noting that "If content properties just are [neurophysiological] properties, there is no reason whatever for thinking content doesn't enter the causal chain leading to behavior" (2011b, 438).[10] And this seems exactly the right thing to say: if the property of *being F* is identical to the property of *being G*, nothing could be caused by virtue of one but not the other. If there is a causal difference between neurophysiological states and belief states, given that causation is a genuine, worldly relation, then that causal difference, via Leibniz Law, would seem to entail *property dualism*, and to rule out type identity theory.[11]

> In fact, in a much earlier response to Jerry Fodor, Plantinga writes the following:
>
> Take first the identity case: the case where the content property just is a combination of neurophysiological properties. Now consider a neural structure S that is a belief, and suppose S displays the combination P of physical properties that is identical with the property of having p as content, for some proposition p. By hypothesis, S's being in P is identical with S's having p as content. Suppose furthermore, that S's being in P is adaptive, fitness enhancing. It follows, of course, that having p as content is fitness enhancing by virtue of the fact that P contributes causally to fitness-enhancing behavior. (2002b, 218)

[10] There is no acknowledgement in the paper that this represents a departure from his position in Chapter 10 of *Where the Conflict Really Lies*. Nor am I sure which came first -- both are published in 2011 -- so I am not certain how the two presentations are supposed to fit together.

[11] No doubt there are possible views that would allow space for what Plantinga is after, perhaps a view that disallows property abstracts of the form "... causes e in virtue of being X". But such views have neither been argued by Plantinga, nor been incorporated into the EAAN.

Again, this seems right. Type identity theory should not have been the starting point for Plantinga's argument in favor of semantic epiphenomenalism.

Could the argument from semantic epiphenomenalism be revived under the assumption of *non*-reductive materialism?

Plantinga could instead have pushed the argument only under *non-reductive materialism*. Recall that he characterizes non-reductive materialism as the view that the mental either supervenes on or "is determined by" the material (2011a, 323). Here, one might argue much more plausibly, as many have, that beliefs are not causal *as such*, and that only their supervenience bases, or determiners, or we might say today, their "grounds", are causal (again, see McLaughlin, 1989).

Semantic epiphenomenalism may very well have a high probability on the assumption of non-reductive materialism, in the final analysis, or from a God's-eye point of view, and therefore, might still be able to provide, *in principle*, some support for Plantinga's claim that true beliefs are not adaptive.

Rhetorically, however, this will be of little use. Philosophy is hard, and these, in particular, are thorny questions, far from settled. No one has settled the objective probability of non-reductive materialism given materialism, nor of semantic epiphenomenalism given non-reductive materialism. Those who assign *reductive* materialism a high probability, conditional on materialism simply won't be bothered by this part of Plantinga's argument. And there are many well-known explanations of how belief-properties in particular and the mental in general could be causally efficacious,

given non-reductive materialism (for an overview, see 6.3-6.4 of Robb and Heil, 2013). Fans of such explanations, again, will think semantic epiphenomenalism unlikely, given non-reductive materialism.

Rhetorically, then, Plantinga would need to respond both to the reductive materialists, and to these many non-reductive yet causal theories, and his responses would need to be compelling. But he has not done so, and I imagine doing so would not be easy, and would instead mire the EAAN in mental causation debates for the foreseeable future.

What Plantinga has given us with (ii) is an argument aptly directed only towards those who think non-reductive materialism is somewhat probable, given materialism, and who also think semantic epiphenomenalism at least somewhat probable given non-reductive materialism. *Those philosophers*, Plantinga might say, should agree that true beliefs are not adaptive, because their contents do not have any causal effects whatsoever.[12] Then *those philosophers* should follow out the rest of the argument to its

[12] To some degree, in his arguments from epiphenomenalism and semantic epiphenomenalism, Plantinga *was* simply taking advantage of existing, though somewhat extreme, positions in the philosophy of mind popular particularly in the 90s. He cites Patricia Churchland (1987) and Steven Stich (1993), both of whom embraced the idea that natural selection does not select for true-belief-producing mechanisms. Churchland famously eliminates beliefs altogether; Stich, though less bold in his conclusions, shares Churchland's skepticism about the utility of belief as taxonomic category, and abandons *true* belief as worthy goal of cognition, in favor of a thoroughgoing pragmatism (Churchland, 1987; Stich, 1993). Plantinga, I think, realized these extreme positions regarding belief run the risk of undermining themselves, as did others, who pointed out that proper assertion, even the assertion of the views in question, requires that the asserter believe what she asserts (Baker, 1987; Boghossian, 1990; Reppert, 1992). But Plantinga narrowed his critique to evolution and naturalism, rather than finding absurdity in assertion generally. For if evolution does not "care about" true belief, then perhaps one should not trust one's own beliefs at all, including of course one's beliefs about naturalism itself, since they are, according to one's own view of things, produced by faculties that result from this truth-indifferent process of natural selection? What is interesting is that the broader arguments that these radical

skeptical conclusion. But *those philosophers*, who I imagine are a very small minority, are not the intended target of the EAAN. Rather, the target is *all* naturalists who adhere to contemporary evolutionary theory. To the extent that the EAAN hinges on arguments for non-reductive vs. reductive materialism, and arguments for semantic epiphenomenalism, given non-reductive materialism, it will not be of broad interest to philosophers, scientists, or the public at large.[13]

Even if beliefs are causal in virtue of their contents, true belief is no more adaptive than false belief, given N&E

I will hereafter drop (ii). We have already dropped (i), so in support of the contention that R is improbable given N&E, that leaves us with:

> (iii): even if beliefs *do* cause behaviors in virtue of their contents, still, *true* beliefs are no more likely to confer selective advantage than *false* ones.

Originally, Plantinga supported (iii) by noting that different belief-desire combinations can lead to the same behavior. He gave examples indicating that there were many possible combinations and circumstances in which false beliefs are as adaptive as true ones. Someone who wants desperately to be eaten, for instance, but who thinks tigers are unlikely to eat him and therefore avoids tigers, will survive tiger threats as well as someone who doesn't want to be eaten and thinks tigers will eat him (Plantinga, 1992, 225-226). Moreover, this sort of adaptive false belief could be global and not merely

---

views are self-undermining did not blossom into a literature, but mostly withered away in the 90s, while Plantinga's narrowing of those arguments lived on.

[13] Not only is the EAAN the final chapter of Plantinga's popularly-oriented book (2011a), but he delivers it as a popular lecture to non-specialists (2013).

local. Suppose a devout animist prefixes "that conscious thing …" to *every* belief of hers about particular non-conscious things (Plantinga, 2002a, 9). This sort of animist will presumably fare as well in an environment as a non-animist, yet every one of her beliefs about non-conscious things will be false. Let us call the argument for (iii) by way of spelling out various false-yet-adaptive belief-desire combinations the *appeal to scenarios*.

The *appeal to scenarios* does not appear in Plantinga's recent presentations of the EAAN, though it is not clear exactly which criticisms, if any, prompted him to drop it. Perhaps it was the objection that one knows what it feels like to want to be eaten by a predator, and to think that the best way to be eaten is to run away from that predator, and consequently, to suffer repeated disappointment at not being eaten. Since that is not what it feels like, *for us*, when we face a predator, we needn't worry about about that sort of perverse assignment of belief content *for us*. Likewise, our beliefs do not seem to have a *conjunctive structure* where conjunction elimination is somehow impossible, so we can never screen off our animism, or other false but stubbornly conjoined beliefs. (How the bookkeeping should go for this sort of objection -- is it to premise 1, the probability premise, or to premise 2, the defeat premise? -- is tricky, as we will see.) Or maybe, Plantinga is worried about objections like Stephen Law's that under an open-ended set of future conditions, say, and given a systematic assignment of belief contents, one can show that such perverse belief assignments will have maladaptive consequences (Law, 2011). Maybe neither of these responses moved him, but some other response, for instance, that while these scenarios are possible, they are not equally probable, or maybe he just saw that the scenarios were not working on his audiences and he didn't need them

anyway. All that is known is that Plantinga no longer appeals to the scenarios to make his case.

Instead, he now relies on another line of reasoning, which we might call the *argument from semantic silence*. The argument here is that, in general, knowing that a particular neurophysiological state is a *belief* that is *adaptive* (even adaptive qua *belief that P,* say) tells one absolutely nothing, one way or the other, about whether that belief is also *true*. So there is simply no reason to think that selecting for adaptivity selects for true belief as well.

Again, Plantinga asks us to think about hypothetical creatures much more primitive than ourselves, but who form beliefs. We are to consider a random belief of such a creature and ask ourselves what we know about that creature:

> The fact that these creatures have survived and evolved, that their cognitive equipment was good enough to enable their ancestors to survive and reproduce--that fact would tell us nothing at all about the *truth* of their beliefs or the reliability of their cognitive faculties. It would tell us something about the neurophysiological properties of a given belief; it would tell us that by virtue of these properties, that belief has played a role in the production of adaptive behavior. But it would tell us nothing about the truth of the content of that belief: its content might be true, but might with equal probability be false (2011a, 331)

In other words, the adaptivity of a belief is simply silent on its truth. On N&E alone, the only assumption about ourselves to which we are entitled is that it is likely that our

beliefs are adaptive. If adaptivity does not make truth probable, we cannot infer that our beliefs are likely true, on N&E alone.

Indicators versus beliefs

Plantinga notes that the needs of organisms may be well served by having "indicators", which are states that nomically or causally or counterfactually covary with certain external states (2011a, 328). Those indicators will be crucial to adaptive behavior: finding food, avoiding predators, reproducing, and the like. But indicators need not be beliefs: thermostats have indicators, but not beliefs. Moreover, the belief states of creatures may have nothing whatsoever to do with their indicators. If thermostats did have beliefs, even beliefs about the temperature, those beliefs might have nothing to do with their internal states that correlate with temperature. There is just no reason to think that the determination of the content of *beliefs* goes in the way we might think it does, i.e., that a creature that is tracking information in its environment and using that information to survive is also forming true beliefs corresponding to what it is tracking.

Plantinga is even willing to concede that beliefs may indicate as well, i.e., that the neurophysiological state that is the belief that P may indicate a state of the world, Q, and thereby correlate with that state. But even if beliefs are indicators, he says, "We know of no reason why the content of a belief should match what that belief (together perhaps, with other structures) indicates" (2011a, 331). A given neurophysiological state might be a belief, and might indicate, for example, the presence of a tiger: it might be instanced when, and only when, a tiger is around, and it may guide the behavior of the organism in

a way that preserves its life by avoiding the tiger. Still, says Plantinga, we know nothing of which proposition that belief (which is also an indicator) is a belief *of*: "Indeed, the proposition constituting that content need not be so much as about that predator; it certainly need not be true" (2011a, 331).

The Equiprobability of Arbitrary Semantic Assignments

And here, though he does not quite come out and say it, is the core of the argument from semantic silence. When Plantinga suggests that a belief that is indicating one thing may have totally unrelated content, he is envisioning scenarios that I think most of his audience is likely ignoring, which go well beyond his earlier *appeal to scenarios*. Let's fill in the argument just a bit. Plantinga is claiming that belief contents and indicator contents (what a neurophysiological state indicates) are, at least in some sense, *totally unrelated*. Think about the neurophysiological state of an organism that is tied to tiger presence, and which triggers flight behavior protocols, say, or a sequence that first determines whether hiding, fleeing, or fighting would be best, primes the organism for action, and then does one of the three, and which is triggered by visual impressions of tigers, or sounds of tigers, or even of sounds of people warning about imminent tiger attacks in natural language. Suppose that neurophysiological state happens also to be a belief. Plantinga thinks there is simply nothing more probable about that state being a belief *having to do with tigers* than a belief *not having to do with tigers at all*. Recall that

"... the proposition… need not be so much as about that predator" (2011a, 331).[14] For all we have said, this tiger-correlative state might be the belief that *it is raining*, or that *unicorns exist*, or that *the European Union's days are numbered*. Given that the content of this particular adaptive neurophysiological state that helps one avoid tigers is totally unknown, and given that the content could with equal probability be *absolutely any proposition whatsoever*, why think it is more likely true than not? This, we might call the thesis of *equiprobability of semantic assignments*, and it underlies the argument from semantic silence.

This argument is much stronger than the *appeal to cases*, where, for instance, tiger indicators that were beliefs were *still* in some way *about tigers*, or beliefs in general were *about* what they indicated but supplemented with false conjuncts. Those examples worked by *adding junk*, so to speak -- strange coordinate desires or auxiliary beliefs, or inert and false conjuncts appended to every belief. One might have appealed to simplicity in favor of an assignment that cut out the junk. Here, by contrast the divorce between indication and content is total. One does not add "junk" to what is the intuitively assigned content of a belief state, namely, the indicator content, to make false beliefs as adaptive as true beliefs. Rather, one discards the intuitively assigned content altogether and chooses a proposition at random.

Return to our hypothetical population, not necessarily ourselves, that evolved in a godless materialistic universe. Now think about the probability that their beliefs are true

[14] See also (Plantinga 2002b, 259): "... natural selection will not be able to select for mechanisms that produce *inaccurate* [indicator] representations. But none of this, so far, has anything to do with *belief*, or with the *truth* of belief."

if semantic assignments to neurophysiological states really are completely *arbitrary*, given materialism, and have nothing whatsoever to do with indication. It is hard to see why having true beliefs would enhance fitness for our hypothetical population. To determine whether true beliefs would enhance fitness, we would need to think about how true believers would behave differently from false believers. But since belief assignments are randomly related to indicators, or perhaps more accurately, simply not related in any discernable way at all, we have no idea at all how truth believers, as opposed to falsehood believers, *would* behave in various environments.

It helps to bear in mind the most extreme divorce between indicator and belief contents: *negation*. The neurophysiological state that *indicates* that *p* (say, that predators lurk nearby) could be the neurophysiological state that has as belief content, *not-p* (that there are *not* predators lurking nearby). If the indicator that p is identical to the belief that not-p, a false believer will do much better than a true one, all else being equal. Imagine a creature that believes not-p if and only if p is true, and whose belief that p is identical to the indicator that not-p. Such a creature will act just like a creature who has true beliefs about the presence of predators under the "standard" kind of assignment you, naive philosopher, had in mind, where the belief and indicator contents coincide. Therefore, under this possible assignment, false belief will be more adaptive than true belief.

So this is how the argument goes, or could go at any rate: true beliefs are not adaptive because beliefs are just neurophysiological states and the assignment of belief contents to those states is utterly arbitrary, from the perspective of someone who knows only about the material world anyway. Any argument that a given true belief is more

adaptive than false belief would have been must suppose that some particular assignment of content is somehow *better* than the assignment of the *negation* of that very same content to the same belief state. For the inverted assignment of content would yield, by hypothesis, creatures that do vastly better with false beliefs than with true ones. But the supposition that one assignment is *better* than another is groundless, given materialism. Every belief content assignment for a given neurophysiological state is as good a candidate as every other. Therefore, there can be no grounds for thinking that true belief is adaptive.

Yet one more way of putting the argument: "true beliefs are adaptive" expresses something, but as a materialist, one really has no idea *what one is saying*, in saying it. Let's call the function that maps neurophysiological states to their contents "*I*" for *interpretation function*. Given that mappings from neurophysiological states to propositions are all equiprobable, on materialism, we have no idea what *I* is. When we say "true beliefs", that is equivalent to saying, "neurophysiological states mapped by *I* to true propositions." Since we have no idea what *I* is, we also have no idea what states we are talking about. Since we have no idea which states we are talking about, we have no idea if *those* states are adaptive or not.

To make the point once again in terms of negation, but now generally, suppose one is tempted by a conjecture, C*, according to which true beliefs are in fact adaptive. C* will employ, tacitly or overtly, some candidate interpretation function, *I**, about the contents of belief states. But now consider another conjecture, C**, which supposes an interpretation function, *I*** that maps every state to the negation of the proposition to

which *I** maps it. Since interpretation functions are all equiprobable, *I*** is as likely as *I**, and therefore C**, insofar as we have described it, is as likely as C*, and it is as likely that false belief is adaptive as that true belief is adaptive.

That is the argument in its strongest form, never stated outright by Plantinga but implied by his remarks that indicator content tells us nothing about belief content (2002b, 259). Notice that the *appeal to scenarios* is still live: nothing in the equiprobability of semantic assignments conflicts with it. And notice that beliefs may still cause behavior, since these beliefs are, by hypothesis, causing adaptive behavior. Under the assumption of reductive materialism, it is the *contents* of the beliefs, or the beliefs-qua-beliefs that are doing the causing too. For now, we can invoke Leibniz's Law, as we did before. If belief types are identical to neurophysiological state types, then if those neurophysiological states are, qua neurophysiological states, causing adaptive behavior, then so are the beliefs. They are one and the same thing, after all. But we don't know, in a given case of an adaptive belief, whether it is *false* contents or *true* ones that are causing the adaptive behavior. From our perspective, or even from the perspective of an ideal reasoner with complete knowledge of the material, these possibilities are equally likely.

Is Plantinga *really* assuming *equi*probability? Two kinds of arbitrariness.

It may seem that I have read too much into Plantinga, foisting on him the argument from equiprobability of arbitrary semantic assignments, which may strike the reader as outlandish and unfair, when what he says is simply that indicator content need not be at

all related to semantic content: "... there is no reason why that content need be related to what the structure indicates, if anything" (Plantinga, 2011a, 331). He does not explicitly say that there are *absolutely no constraints on content whatsoever*. He does not use my argument from the equiprobability of inverted assignments. (He should!)

But if he does envision any constraints, he never gives his reader any reason to think so. Nor I am not alone in my reading. Stephen Law, at least, shares it (2012).[15] In a response to Plantinga, Law argues that there are certain "conceptual constraints on content" that link belief and behavior and thereby favor some possible content assignments over others. Law remarks that "If such constraints exist, then one cannot, as it were, plug any old belief content into any old neural structure, irrespective of that structure's behavioral output" (2012, 6). Law thereby implies that Plantinga thinks one *can* in fact "plug any old belief content into any old neural structure".

Law repeats this point multiple times in a joint podcast with Plantinga, using the exact phrases "any old belief" and "any old neural structure" (*Plantinga and Law, 2010*). Plantinga does not dispute the characterization *per se*, but responds with two qualifying points: first, *actually* yes of course true beliefs are *in fact* adaptive, and actually adaptive beliefs are mostly true. But we are not talking about what is actually the case, but what *would* be the case given materialism. Under the assumption of materialism, we really

[15] If Law and I are both wrong, and Plantinga does not actually mean to say that content assignments are *truly* arbitrary, then it is not clear exactly how Plantinga's argument that truth belief is no more adaptive than false belief is supposed to go. If he falls back on the *appeal to scenarios*, Law actually has a plausible response (2011). Law argues that perverse combinations of beliefs and desires will, across a wide range of scenarios, prove maladaptive, given open-ended interactions with new beliefs and desires in a belief system in a changing environment (2011). If it is is some middle ground, between arbitrariness and the scenarios, one wonders whether the non-arbitrariness might actually somehow favor true beliefs. We have no way of knowing, since this middle way has not been developed.

have no more reason to think true belief contents are assigned to adaptive states than false ones. Secondly, if materialism, either reductive or non-reductive materialism, is true, belief assignments are constrained by identity or supervenience relations, and therefore in *some* sense *non*-arbitrary, yet still no more likely true than false (*Plantinga and Law, 2010*).

Both of those responses, I think, are consistent with my characterization of the argument, and with Law's. The first point does not even pose a prima facie difficulty, though I will return to it for other reasons. The second response -- that content assignments are *not* arbitrary, given the supervenience or identity relations between contents and neural states does not pose a problem either, though it may seem to at first. It actually helps to illustrate the nature of the probability judgment in question.

Suppose that something can have non-zero probability *only if* it is metaphysically possible. Then, on type identity theory, all but one of the infinitely many assignments of neural states to belief contents will have a probability of *zero*. Since identity is necessary, the very same neural state types, or (neural-state-types-in-environment-types) are mapped to the same contents in every possible world by the identity relation.[16] Thus, by our supposed link between metaphysical possibility and probability, the probability of the mapping going any way other than the way it actually does is *zero*.

[16] Of course, given content externalism, the mapping will somehow have to include information about the environment, both generic and particular, to accommodate beliefs about kinds and about particular individuals. Though this raises complicated issues about how the belief is only identical to the neural state, when the world plays a role in determining content, I will ignore them for the purposes of this paper. Assume, to make it easy, that neural state types include information about the environment (e.g., such-and-such arrangement of neurons related thus-and-so to H_2O and thus-and-so to Bob).

So probability, in the sense that Plantinga intends it, cannot have the link to possibility that we have just supposed. For then, we should not confidently assign a probability of "low" to (R|N&E). For it might be on that modal understanding of probability, that the probability of R is actually zero, or very very high. Suppose that naturalism is true, and that every neural state is mapped to exactly the same proposition: the proposition that 1=0. If these mappings are necessary, then necessarily, every belief of every organism is false. So the objective probability of R, on N&E, is *zero*. On the other hand, maybe every state is mapped to the proposition that 0=0, and the probability of R is *one*. Or, we might suppose that belief contents are actually very closely tied to indicator contents, in which case they would be *necessarily* so tied. And since indicators are, by hypothesis, correlative with what they indicate, these beliefs would stand a good chance of being true. One should not represent this complex state of affairs -- that the probability of R|N&E could be zero, low, one, or very high, by saying simply that the probability of R|N&E *is* low.

Of course, it may be that *we have no reason to think* that indicator contents are tied to belief contents in any particular way, no reason to think one identity relation, rather than another, holds, between belief contents and neural states . But if that is the argument, then the first premise should not be *P(R|N&E) is low*, but rather, that *we have no good reason to believe that P(R|N&E) is anything but low*. For all we know, or even for all an ideal reasoner who knew all of the material facts would know, the semantic assignments to neural states may go any which way. But if the notion of probability is really meant to track *metaphysical possibility*, it would be presumptuous to declare

straight away and without qualification that the probability of R is low on N&E. It would be like a pre-Babylonian astronomer, one who is already on board with the necessity of identity, and who is aware that "Hesperus" and "Phosphorus" are rigid designators, confidently declaring that it is metaphysically possible that Hesperus isn't Phosphorus. It would be better for the astronomer to say that *for all we know, it is possible that Hesperus isn't Phosphorus*.

The problem for Plantinga is that he is saying two very different things about the relation between indicator and belief contents. In the *first* place, he says the one kind of content may have nothing to do with the other, and it is the equiprobability of alternative ways this mapping could have gone that forces us to agnosticism about whether a particular adaptive state that happens to be a belief is also true: the belief is just as probably false as true because there are equiprobable semantic assignments of belief contents that assign the state to true and to false propositions. But in the *second* place, he argues that the semantic assignments are *not* arbitrary, because, on the assumption of materialism, they are *metaphysically necessary*. (They're constrained by either identity or supervenience relations.) These two things are prima facie in tension. Either the assignments are arbitrary or they are not.

Exactly how he resolves things does not matter to me. One way would be to render objective probability in terms of the credence an ideal reasoner would adopt, where that ideal reasoner is suitable constrained in her knowledge by what is being conditionalized upon. Ignorant of the identities in question, or the supervenience relations in question, an ideal reasoner is still stuck with equiprobability for all

conceivable assignments of content (according to the argument). It is not clear to me that this is mere epistemic probability either, since the reasoner is idealized. We may think of our subject matter -- objective probability -- as the categorical basis for the disposition to bring about a certain degree of credence in an ideal reasoner under certain noetic constraints. That categorical basis is not epistemic even though its manifestations are.

Another way to resolve Plantinga's tension would be to leave the connection between objective probability and metaphysical possibility in place and to replace premise 1 with the claim that *even an ideal reasoner with full knowledge of the material would have no good reason to believe that P(R|N&E) is anything but low*. That would also suffice. What is important is that there are two senses of "arbitrary" in play here, only one of which is tied to metaphysical possibility. [17]

The inadmissibility of centering information

[17] In fact, matters are even more complicated. If Plantinga is right and materialism is false, and necessarily so, then there are no identity or supervenience relations between material states and contents. So then, on the one hand, the indicator content-semantic content relation is not *constrained* by metaphysical relations, and seems, from that perspective, truly arbitrary. But given the necessity of identity, it is also impossible that any indicator *could* be identical to a belief, so, if we are thinking about a metaphysical notion of probability, and assuming Plantinga is right that materialism is false, *none* of the mappings the materialist has in mind has a non-zero probability. Beliefs are, and necessarily are, *not* identical to neural states. The point, once again, is that in the context of a reductio-style argument, which the EAAN is, one cannot tie objective probability to metaphysical possibility, since the latter hinges very much on what is actual (actual identities, for instance). Instead, one needs a much more flexible notion, one that allows non-zero, non-one, probabilities of metaphysical impossibilities and necessities, respectively. One needs probabilities that are conditional on an impossibility -- materialism. One needs a much more flexible notion, and it is on that *flexible* notion that different assignments are equiprobable (arbitrary), and it is on an inflexible notion, tied to metaphysical possibility and tied to the actual, on which semantic assignments are not equiprobable (not arbitrary).

Now let's return to the first point, that what matters is not whether true beliefs are *actually* adaptive -- he concedes that they are -- but whether they would be, if naturalism were true. Here again, we see that the notion used in the probability premise is not a straightforward modal notion. Plantinga considers the actual adaptivity of true belief: "A gazelle who mistakenly believes that lions are friendly, overgrown house cats won't be long for this world" (2011a, 335). But this observation is "irrelevant" to the argument, for what matters is "… what things would be like if N&E were true; and in this context we can't just assume, of course, that if N&E, N including materialism, were true, then things would still be the way they are" (2011a, 336). Plantinga believes that if materialism *were* true, even reductive materialism, true belief *would not* be adaptive (or more adaptive than false belief).

But if one believes both reductive materialism and that true belief is adaptive, and if one subscribes to the standard semantics for counterfactuals on which if the antecedent and consequent of a counterfactual are both true then the counterfactual is true, then why *wouldn't* one simply say, "If reductive materialism *were* true, true belief would *still* be adaptive, because reductive materialism *is* true and true belief *is* adaptive"? That's like asking, on a warm sunny day, whether one would be comfortable if it were sunny. Supposing one actually is comfortable, the answer is a simple "yes". It is not clear why the appeal to actuality is off limits. The explanation, I think, is that Plantinga is using counterfactuals here as an imperfect proxy for objective probability, and such "centering" information is not "admissible" when considering the *probability* of R on N&E *alone*. So

Plantinga's invocation of the counterfactual is in fact somewhat misleading.[18] It is only the counterfactual considered in ignorance of information about the actual -- the sort that an ideal reasoner would not have -- that roughly corresponds to Plantinga's notion of conditional probability. And again, on that notion, all belief content assignments will be equiprobable.

How to debate the probability premise?

As discussed in the previous section, Plantinga thinks that on the assumption of materialism, if we choose a random adaptive state of an organism, a state that also happens to be a *belief*, the adaptivity of that state tells us nothing whatsoever about the *contents* of that belief -- it could be a belief about anything-- and therefore, tells us nothing about the *truth* of that belief -- that random, unknown, content proposition might be true or might be false. I will not try to survey the objections to this premise, though we will delve into some of them later on. Rather, I want to try to illustrate, in this section, the difficulty involved in actually objecting to the premise, and the way in which

[18] Looking ahead a bit, the disconnect here between the counterfactual and the conditional probability may also be revelatory. For in the "defeat" premise, we are told that this objective probability belief defeats our belief in R, and hence all of our beliefs. If the objective probability claim were indeed a *counterfactual*, the "defeat" would be far more apparent. Consider: "I believe if N&E *were* the case, R *would* be highly improbable, and I also believe that N&E *are* the case, and that R *is* in fact true." Very odd sounding! That sounds like saying "If it were sunny, I'd be comfortable. It is sunny, but I'm not comfortable." That sounds very very bad.

So perhaps some of the support for the "defeat" premise, in fact, derives from the plausibility of this counterfactual version and not the objective probability claim, which is really quite distinct from the counterfactual. But the counterfactual version renders the "probability premise" false or at least unpersuasive for a naturalist, as I argued above. For the naturalist thinks N&E are *actually* true, along with R. And therefore, that R *would* be true, given N&E.

objections to the premise draw us into the gravitational field of the second premise, the defeat premise, instead.

Stephen Law objected, as noted earlier, that perhaps there are conceptual constraints tying beliefs to behavior in characteristic ways that make adaptive beliefs more likely to be true than false (2012). Law asks us to imagine a thirsty human with a desire for water, who will survive only if he walks five miles south to the nearest available water, and who in fact makes the southerly trek, drinks the water, and survives (2012, 5). Law thinks that there are conceptual constraints on belief content assignments such that, assuming there is a neural state of our subject that characteristically causes in its possessors such southerly walks under conditions of dire thirst, followed by bouts of drinking that water, it is likely that our subject believes, prior to his journey, that there is water to the south (2012, 5). What exactly the constraints are, Law does not say, but they underlie our judgment in this case and cases like it, that having true rather than false beliefs is fitness enhancing.

Plantinga's reply is that he just *does not see* these conceptual constraints on belief: he does not think it likely at all, on materialism, that the subject in question *does* have a true belief about the whereabouts of water, rather than some other belief that happens to be false (*Plantinga and Law, 2010*). Of course, the subject may have a state that *indicates* the presence of water to the south. But that indicator need have no relation to the subject's belief (*Plantinga and Law, 2010*).

Recently, Calum Miller has taken up Plantinga's cause against Law (2015). Miller labels Law's proposal that there are conceptual constraints on belief that make

adaptive beliefs more likely true than false "CC+" (2015, 149). Miller argues that CC+ is itself improbable, given N&E, for CC+ amounts, roughly, to R itself, and Plantinga's whole point is that R is improbable, given N&E, and therefore, so is CC+ (2015, 147, 149). Miller thinks Law is begging the question.

It is hard to know how this debate should be settled. By the lights of Plantinga and Miller, Law is simply not taking the presumption of materialism seriously enough. If he did, he would have to give up on his proposed connections between belief contents, successful behavior, and truth. As it is, he is begging the question. But by Law's own lights, he is taking materialism quite seriously and still he sees such connections, and is *not* begging the question.

The issues here are extremely subtle, and cannot be probed without entering into the debates over premise 2, the defeat premise. For, one way of presenting Law's proposal is as an argument against premise 1: Law is arguing that the probability of R given N&E is *not* low. But another way of making Law's point is that the probability of R, given only N&E is indeed low, but the probability of R given N&E*&CC+* is high. (This is Miller's reading of Law.) And on this second way of going, Law would be denying the defeat premise, not the probability premise. He would be allowing that P(R|N&E) is low, but arguing that acknowledging that probability does not make belief in R irrational, since we also believe CC+. Whether such a move is *permissible* is a question, in Plantinga's terminology, of whether CC+ is an "admissible deflector", and one can interpret Miller as arguing that CC+ is *not* an admissible deflector.

I will discuss deflectors at length in the following section, on the defeat premise. For now, simply note that a "deflector" for this particular defeater is a further belief, Q such that P(R|N&E&Q) is not low and Q does not beg the question, so to speak, against the putative defeater. On Plantinga and Miller's construal of Law's objection, then, CC+ is offered as a deflector, and it is not admissible because it begs the question.

I think Law should stick to his guns and argue that if the constraints on belief assignments are truly *conceptual* constraints, then given the right notion of objective probability, the probability of R on N&E together with these constraints should equal the probability of R|N&E. An ideal reasoner, for instance, would assign the negation of the constraints probability zero. (Compare: P(R|N&E) should equal the P(R|N&E&(x)x=x).) Therefore, he is not offering a *defeater deflector*, but simply objecting to premise 1 in the first place.[19]

But obviously, Plantinga and Miller do not think there are any such conceptual constraints. And they think the only reason Law thinks there are is that we have a "plausible and widespread belief" (Miller, 149) that true belief is adaptive. Thus, from their perspective, Law is offering a candidate "defeater deflector", which must avoid question begging, and which they think does *not* avoid question begging.

In "Content and Natural Selection", Plantinga canvasses several naturalized theories of content: Fred Dretske's indicator semantics, functionalism, and Ruth

---

[19] There is a little wrinkle here. What Miller labels "CC+" are not actually the conceptual constraints themselves, but the proposition *that such constraints exist*. But if there are such constraints, and if they get probability 1 from an ideal reasoner, the proposition *that* they exist should also get probability 1 from an ideal reasoner. Compare: if Leibniz's Law is a conceptual constraint on identity, then it is a conceptual truth *that a constraint exists* limiting the properties that x has, given that x=y.

Millikan's teleosemantics (2011b). All of these theories, if true, would link up true belief and adaptive behavior in some fashion. One could then see these theories as attacks on the probability premise, for if they are true, then P(R|N&E) is not low after all. But that is *not* Plantinga's approach. Instead, he considers each theory as a potential "defeater-deflector", and argues that, for one reason or another, none is admissible as a deflector. So he takes them to be possible attacks on the *defeat* premise, not the *probability* premise.

Again, it is not clear that this is the right way to construe these theories. It may be that we should consider them as arguments against Plantinga's original probability judgment and not as palliatives to it that must meet Plantinga's standards for not begging the question against him. But that question can be responsibly addressed only after we consider the defeat premise itself.

Further difficulties in assigning objections to premises

I want to make one more remark before moving on to the defeat premise itself, because there is a subtlety here that cannot be ignored. It may be that in estimating the probability of R on N&E, Law *is*, in fact, somehow *influenced* in his judgment by his prior beliefs in the actual truth of N (or something close to it), E, and R. Likewise for the various naturalizers of content surveyed by Plantinga. But this influence could take different forms. It could be that Law really introduces some *supplemental* proposition in addition to the probability premise, which stands alone, and thereby, in Plantinga's eyes, begging

the question. Or it could be that he is simply *more confident in the rationality of his belief in R*, than he is in an evaluation of the probability of R|N&E.

A helpful comparison here is a possible position one could take on the relation between freedom and determinism. Suppose someone is an incompatibilist, and therefore believes that he cannot be both free and determined, and therefore, that if determinism *were* true, then we *wouldn't* be free. But suppose he also admits that if he happens to discover that determinism is *actually* true, from an unimpeachable source, then he will abandon his incompatibilism rather than deny that he is free.[20] Or think of a dualist who believes that if she *were* a minimal physical duplicate of herself, she would not be conscious -- her possible minimal physical duplicates are zombies -- but who, if she is told by an oracle that she is, *in fact*, actually a minimal physical duplicate of herself, would abandon her earlier belief in the counterfactual, abandon her dualism, and continue to believe that she is conscious, rather than follow through on her previous belief, and conclude that she himself is a zombie.[21]

In fact, one wonders what Plantinga himself would do, were he to become utterly convinced of the truth of naturalism and evolution. Would he abandon R, or continue to believe, in any case, that abandoning R is the *right* thing to do, even if he cannot quite

---

[20] (Here, I want to cite a friend who holds this view, not in any publications, but have yet to ask his permission).

[21] John Hawthorne, in "Advice for Physicalists" writes:

> Well, suppose an oracle tells you tomorrow that the world is merely physical. Will you conclude that there is no pain, that your earlier self was making a mistake in ascribing pain to himself on occasion? No. You will remain convinced that you do feel pain sometimes and will reckon as pain whatever plays the pain role. (Relatedly, you will form the belief that being conscious of that state is not a non-physical, unanalyzable acquaintance relation, but instead some sort of causal/functional relationship to the state.) (2002, 26)

muster the courage? Or would he think instead, "I must have gone wrong somewhere in my argument. Probably premise 1"? It does not seem obvious to me that the latter response is irrational. And it seems very different from thinking, "I must have gone wrong somewhere, *and therefore must have a defeater deflector*." But I am also not sure what to make of this strategy, particularly in this instance. It does feel like the theoretician, here, is not following through on her initial judgments, when she *first* believes that if p were the case, q would be the case, *then* discovers p and changes her mind about the counterfactual, rather than believing q.

But why exactly is it wrong, epistemically speaking, to change one's mind about the counterfactual upon receiving new information about the actual world? Maybe one is *highly uncertain* of the counterfactual, but *extremely certain* of its consequent. In that case, when a conflict emerges, why is one stuck with the previous belief in which one was *least* certain? Wouldn't the sound, Moorean advice be *against* sticking to the *less* certain belief in cases of conflicting beliefs and *for* sticking to the *more* certain belief? And in this case, figuring out how indicator and belief contents are connected seems very hard; the issues surrounding it are highly abstract and we are not so great at reasoning about such highly abstract matters. But *clearly*, true belief is adaptive. That is just *obvious*, one might think. And for all the EAAN says, we may also have excellent reasons in favor of naturalism and evolution, much better reasons than we have for thinking that belief and indicator content are not connected. So if there is a tension, why *shouldn't* the tension find release in a shift in the naturalist's initial judgment of the probability premise itself?

Perhaps this is just one manifestation of a deeper epistemic truth, which is that one of the "two dogmas" of empiricism is false, and that confirmation relations are holistic. As Quine says, "our statements about the external world face the tribunal of sense experience not individually, but only as a corporate body" (Quine, 41). It is not clear that we can ever really *isolate* P(R|N&E) from the rest of our beliefs. If we could isolate P(R|N&E), then Plantinga, Miller, and critics like Law should theoretically converge in their probability estimates, or at least their disagreement about the actual truth of naturalism would in no way bear on those estimates. But it seems they cannot, or have not anyway, converged. Perhaps it is impossible to *isolate* this premise, even though it is a judgment of objective probability, and even if truths about objective probability have something like the status of logic. On confirmation holism, or epistemic holism, convergence is not to be expected, even here. We are simply not ideal reasoners with perfect access to objective probabilities, independent of all of our other beliefs. Our access to objective probabilities may always be colored by our beliefs as a whole.

Another deep problem here, closely related, and again inspired by Quine, is that it seems that Law's criticism, and others like it, which hold that the probability of (R|N&E) is not low after all, should be cast as criticisms of premise 1 *only if they are conceptual truths*, for only then will the probability of R given N&E equal the probability of R given N&E together with the reasons for doubting premise 1, in which case these reasons are not really *separate* addenda. If the reasons for doubting premise 1 are *not* conceptual truths, then these probabilities will *not* be *equal*, and it seems the reasons should then be cast as defeater deflectors, and therefore as criticisms of premise 2 instead. But it is

unclear what someone like Quine, or anyone who follows him in denying the legitimacy of the category of *conceptual truth* to begin with is supposed to do. It seems they *cannot* actually criticize premise 1, by Plantinga and Miller's rules of engagement. They might simply reject it, without offering a reason. But as soon as they offer a *reason* to doubt premise 1, that reason will be considered a "defeater deflector" and evaluated as such, thereby pulling them into Plantinga's tangle of "admissibility requirements" to be discussed shortly.

Probability Premise Summary

We have waded into deep epistemic waters. I don't know how to resolve the worries I have raised. I wish the dialectic were clearer and more straightforward, and that it did not involve us in these sorts of meta questions about whether one must follow through on the counterfactuals one believes upon discovery of their antecedents, or whether Plantinga really gets to recast any critic of premise 1, or at least any such critic who does not promise a "conceptual" criticism, as someone actually criticizing premise 2, and who is thereby subject to the rules for evaluating defeater deflectors, which, as we will see shortly, are terribly unclear. I wish we did not have to mire debates about premise 1 in the further questions of whether there *are* conceptual truths, or whether confirmation is really holistic. But that is, in fact, where the dialectic goes.

For now, let me recap the discussion of the probability premise. Plantinga claims that the adaptivity of a belief gives us no reason to think the belief is likely true. Hence, since natural selection operates only by filtering out maladaptive traits, we have no

reason to think that the results of an unguided process of natural selection have mostly true beliefs. Therefore, we have no reason to think believing creatures in general are likely to have true beliefs, or faculties that produce such beliefs (reliable faculties), given N&E. But then, we should also assign the same low probability to the proposition that we ourselves have reliable facilities, given N&E. Thus, *P(R|N&E) is low*.

The argument that nature does not select *against* false beliefs, or *for* true ones, varies, but I set aside the appeals to epiphenomenalism and semantic epiphenomenalism. Those arguments proceeded by saying that under the assumption of materialism, beliefs are not causal, or not causal *in virtue of their content properties* anyway, and therefore, do not contribute to fitness. But epiphenomenalism holds little appeal for most contemporary materialists, and will not prove rhetorically useful for Plantinga unless he can unleash new and compelling considerations to sway the majority of materialists who think beliefs are causally efficacious. *Semantic* epiphenomenalism, on the other hand, is presented, confusingly, under the heading of reductive materialism, where it can be shown -- as Plantinga himself has done -- to be false by a Leibniz-Law argument. I considered semantic epiphenomenalism under the heading of non-reductive materialism instead, as Plantinga should have, and found it more plausible there, yet still rhetorically ineffective, given the highly controversial nature of the current debates about mental causation.

I also considered what I called Plantinga's *argument from scenarios*, in which odd combinations of beliefs and desires proved false beliefs can sometimes be adaptive. But I dropped this argument since: (a) it does not appear in any of the latest presentations of the

EAAN; and (b) it faces Law's objections that such combinations cannot be as adaptive as more orthodox content assignments, given the prospects of open-ended, unspecified future interactions with one's environment.

I then examined Plantinga's argument from what I called the *equiprobability of arbitrary semantic assignments*. The argument, here, is that adaptive beliefs could, in *some* sense of "could", have *any content whatsoever*. While they may *indicate* conditions of the environment of the organism, the contents of those beliefs may (again, in some sense of "may") have nothing to do with what they indicate. Therefore, since we do not have any idea what the contents of adaptive beliefs are, we have no grounds for saying that adaptive beliefs are likely to have *true* contents. I argued that this is the best way of understanding Plantinga's argument for the probability premise.

Finally, I reviewed Stephen Law's objection to the probability premise. Law claims there are conceptual constraints on belief that link it to behavior in such a way that adaptive beliefs are more likely true than false. Plantinga's response, along with Callum Miller's, is that the existence of such constraints is, in fact, unlikely given N&E. Breaking the Law-Plantinga stanoff on this point was left for a discussion of the defeat premise, since both Miller and Plantinga see Law as relying on more beliefs about actuality than just N&E, and that, they say, is a response to the *defeat* premise, not the *probability* premise. Furthermore, other attempts, like Law's, to dispute premise 1 from the nature of belief are also treated as "defeater deflectors", i.e., as responses to premise 2. It is unclear, actually, whether we must follow Miller and Plantinga's lead here, but that issue itself is best tackled after a discussion of premise 2, to which we now turn.

The Defeat Premise: Anyone who accepts both N&E and that P(R|N&E) is low has a defeater for R.

Plantinga's argument for the defeat premise is couched entirely in the language of "defeat". This is unfortunate for two reasons. First, the language is technical, and introduced and defined in different ways in his different works. The defeat premise is the most disputed premise of the argument, and our intuitive grasp on the content of the premise -- what it is saying -- is weakened by having to translate familiar notions into Plantinga's technical language. Secondly, the work on defeaters and defeat generally has not coalesced into what we might think is a natural kind about whose contours are understood and agreed upon. Quite the opposite. I would not go so far as John Hawthorne, who, speaking of the literature on defeat says: "... much of that work is of such tragically poor quality that it is not at all clear what can be learnt from it" (2007, 10). Nevertheless, I agree that it does not clarify, but if anything obscures Plantinga's argument.[22] Even so, I cannot avoid engaging with defeat, and Plantinga's thoughts on defeat, altogether, for that is how he himself frames the argument. My strategy, then, is to briefly introduce the premise by way of *defeat*, and the accompanying notions such as: *undercutting defeater*, *rebutting defeater*, *defeater-defeater, defeater-deflector*, and so on, but then to try to state the argument without them.

[22] Indeed, Jonathan Kvanvig has argued, persuasively, that the specific notion of defeat needed by Plantinga for the EAAN is not Plantinga's own, but a rival view (Kvanvig).

The language of defeat: mental state defeaters, undercutters and rebutters, defeater defeaters, and defeater deflectors

Plantinga's notion of a defeater is a *mental state* defeater, not a *propositional* defeater. Propositional defeaters are simply truths, whether or not those truths are believed or known by anyone, that serve some purpose in epistemology, typically, playing a role in Gettier-proofing a theory of knowledge. (Plantinga calls propositional defeaters "warrant defeaters", because they stand in the way of knowledge, but not rational belief) (2011a, 166). By contrast, mental state defeaters are not necessarily true, but they must actually be mental states of the putative knower: they are beliefs or experiences, though for our purposes, we will ignore the *non-belief* types of mental state defeaters (Plantinga, 2000, 363).[23] Which beliefs, or belief-like states, are defeaters? Well, first we need to settle on the logical form of a defeat statement. Beliefs defeat *other beliefs*, and they defeat those beliefs *for a particular subject*, and against the background of what Plantinga calls, their "noetic system" (2000, 363). So the form for defeat statements is: *d is a defeater for belief b for subject S* at time *t*. And d defeats b for S precisely when, given S's noetic structure at *t*, she "cannot rationally hold b, given that [she] believe[s] d" (Plantinga, 2000, 361).

Mental state defeaters come in two varieties: *undercutting* and *rebutting*. The distinction between these two kinds of defeaters, in Plantinga's presentation anyway, is not entirely clear. As Plantinga introduces the distinction, a rebutting defeater is one by way of which the subject "learns that the defeated belief is false", whereas an

[23] Actually, as I shall argue shortly, it may even be the belief state that is the lack of a certain belief that may defeat a belief.

undercutting defeater causes one to lose one's reason for holding belief and leads to agnosticism on the original question (2011a, 165).

It is not clear what happens in cases in which one gets good, but not overwhelming counterevidence. It seems to me this is a case of *partial rebutting*, and not undercutting, though Plantinga's treatment, in (2011a) anyway, would seem to classify it as an undercutting defeater. And what to say about such cases, which I classify as partial rebutting, will *matter*, as it happens.[24]

Plantinga's example of a rebutting defeater: S believes there are no cacti in Michigan's Upper Penninsula, but on a hike comes across a fine specimen of prickly pear (and presumably believes it to be a cactus and in the U.P.). S's new prickly pear belief defeats her previous belief about that absence of cacti in the U.P., and in fact gives her a new belief in the negation of her original belief: *there are indeed cacti in the U.P.* (2011a, 165).

Plantinga's example of an undercutting defeater: S believes, on the basis of visual experience, that Paul has just emerged from the house next door. Then S learns that Paul has a visiting twin brother, Peter. S's belief about the visit defeats her earlier belief about Paul, but *without* convincing S that she did *not* see Paul. Rather, S learns that her visual impression does not strongly support the conclusion she drew from it.[25] S's proper response to the undercutting defeater is to suspend judgment about who emerged from the house next door (2011a, 165).

---

[24] Sometimes rebutting defeaters are also referred to as "overriding" defeaters.

[25] Precisely at this point, certain epistemologists object that if one's idea of evidence is, in Williamsonian fashion, just what one knows, then if one saw Paul, one's evidence is one's seeing, which is a form of knowing, and not a mere impression (Baker-Hytch and Benton, 2015).

Now we can state the defeat premise with a little more precision: the belief in the probability premise works, in the first place, as a *rebutting* defeater for R. Though one certainly does not "learn" that R is false -- one couldn't learn such a thing if Plantinga is right, because one would not be rational in holding any beliefs -- one does learn that R is improbable and R's negation is probable, given one's other beliefs. That is a kind of *partial rebutting*, one that does not rebut one's belief in R *all the way to rational belief in the negation*, but which does its defeating work in the rebutting manner, namely, by bolstering confidence in the negation, and not (immediately, anyway) in the undercutting manner, namely, by calling into question the reliability of the *process* by which the belief came about.

What happens next is that the belief, or awareness, that R is unlikely, given N&E, serves as an *undercutting* defeater for R and everything else. One does not, by way of the probability premise, learn that R is false. But one loses enough confidence in R that one thereby acquires an *undercutting* defeater for everything, including R itself. That is, once one cannot rationally affirm R, then one has a reason to doubt the reliability of the source of all of one's beliefs, including R, supposing one were still to try to believe it. So belief in premise 1 ends up both *partially rebutting* R and consequently *undercutting* everything, including R.[26] Belief in premise 1 thus counts as an instance of *both* kinds of defeater for the same belief.

Putative defeaters can themselves be "defeated," or they can be "deflected". The difference between these two kinds of anti-defeat is important. A defeater d, of a belief

[26] Plantinga never quite presents things this way, but I think it is close to the reading of premise 2 by Michael Bergmann (72, ftnt 26).

b, for an agent S is *itself* defeated when S acquires a belief d*[27], such that it is irrational for S to continue to believe d given that she believes d*. We seem to get the definition of a "defeater-defeater" just by applying the definition of defeat twice. Imagine you believe the departmental meeting is at 3pm, but a friend tells you the meeting is at 3:30pm, thereby defeating your original belief. Suppose, now, a second friend tells you the first friend is lying about the meeting time so as to stack the meeting with his allies on a particular issue that is coming to a vote. Your belief in the second friend's testimony defeats your belief in the first friend's testimony, defeating your defeater. Your original belief, that the meeting starts at 3pm, is thus restored to rationality by a defeater-defeater.

On the other hand, the notion of a defeater-*deflector* is a little bit novel. The idea here is that, since defeat happens relative to an agent, and relative to that agent's entire noetic system, some potential defeater, something that *might* be a defeater relative to *some* noetic systems, could be deflected from actually being a defeater for a particular agent, by something else *that* agent believes. This wouldn't be a case of defeater-defeat, but rather, a case in which there was *never any defeat to begin with*.[28] As Plantinga puts it, for the naturalist:

> Perhaps some of those other propositions [the naturalist believes] are such that by virtue of her believing *them* she doesn't get a defeater for R when she believes

[27] A more careful formulation would allow that mental states other than belief can also defeat defeaters, though I will ignore such states for my purposes.

[28] Plantinga says: "[Defeater-defeaters] would require that one first have a defeater D for R, and then acquire another belief that defeats D. A defeater-deflector, on the other hand, prevents D from being a defeater in the first place" (2002b, 224).

> N&E. Perhaps she has a defeater-deflector for the looming defeat of R threatened by *P(R|N&E) is low* and N&E (2011a, 346).

So, while a defeater-defeater defeats an actual defeater, a defeater deflector does not deflect any actual defeater. For, if it were an actual defeater, but deflected, then it both would, and would never have been, a defeater. Rather, a deflector deflects a "looming" defeater, to use Plantinga's language above. Looming defeaters are thus not defeaters for those subjects who have deflectors.[29] And since there is no such thing as a defeater that is *not* relativized to subjects and their noetic systems, we can say simply that looming defeaters are *not* defeaters.

Interlude: some concerns about the very idea of defeater-deflection

The notion of a defeater-deflector is suspect. It seems to me that *anything* is a looming defeater for *anything else*, in the sense that relative to *some* noetic system, absent "deflectors", it *would* be a defeater. Begin with arbitrary beliefs, *p* and *q*. For an agent with the belief *if p, then not-q*, and who is confident in p, p is a *rebutting defeater* for q. That agent simply reasons through modus ponens to reach not-q, then sees a contradiction -- q and not-q -- and realizes that she cannot hold onto q. But then p is a "looming" defeater for q for everyone: there is some noetic system in which p is a defeater for q, namely, the one just described, in which the subject believes that *if p, then not-q*.

If looming defeaters must be deflected in order to rationally be held, then every rational subject that believes q must have an admissible deflector that "saves" that

[29] Note that Plantinga himself, apart from the quote above, never defines "looming" defeaters or what it is for a belief to be "threatened" by a looming defeater.

believer from looming defeat. p could be "deflected" by, say, the belief *it is* not *the case that if p, then not-q*, but it must be deflected by something, and whatever deflects it must be "admissible" by Plantinga's rules. Admissibility is, as we shall see, very complicated, but at its heart is a ban on question-begging. But since p and q were arbitrarily chosen, we have just shown that everything is a "looming" defeater for everything, and rational belief in any proposition only happens by way of defeater deflectors. If deflectors cannot beg the question against what they deflect, that seems to grant the skeptic entirely too much, at least if the conditions on admissibility are strict, for it says, essentially, that *every belief is irrational unless saved from irrationality in non-question-begging fashion*. And why would any non-skeptic ever admit that?

Plantinga might object that it is not enough for a belief to qualify as a looming defeater of another, or to "threaten" another belief for a subject, that it is, in fact, a defeater of that belief in another subject, as I have supposed. But it also cannot be a defeater for all subjects: it is *never* a defeater for those who have deflectors. So it is just unclear what is meant by "looming" defeat, or "threatening" defeat. Consequently, it is unclear what is meant by "deflector", since deflectors are deflectors *of* looming defeaters.

Perhaps what Plantinga has in mind is something like a "normal" noetic system, relative to which these "looming" defeaters would be actual defeaters. Now, we can say that a proposition is a looming defeater for another proposition only if it would defeat that proposition in one of these *normal* noetic systems. If so, then not every belief would be guilty until proven innocent, for it might be, in the argument above, that no normal

noetic system would contain the belief if p, then not-q, for certain p's and q's. (Let p=q, for instance.)

But if Plantinga says "looming defeat" is not a universal relation holding between all belief pairs, and he takes my suggestion, relativizing to something like "normal" noetic systems, then it is *his burden* to classify the belief *P(R|N&E) is low*, for a naturalist, as a looming defeater in need of deflection. For now, that classification requires the further proposition that in a *normal* noetic system, or *ceteris paribus*, it *would* be a defeater, even for a subject who has lots of beliefs about belief, e.g., that adaptive beliefs are more likely true than false. And I am not sure how that argument would go, exactly, without begging the question against the naturalist, who thinks her noetic system is very much normal, and who goes on believing R, even if she finds premise 1 compelling, and who admits that, by stringent rules of question begging, the auxiliary beliefs about beliefs would be "begging the question" against a skeptic about R, and thereby might not qualify as "deflectors" by Plantinga's criteria. Their argument would be: *P(R|N&E) is low* is not a *looming defeater,* and therefore not in need of *deflection* at all.

The impossibility of Defeater-Defeat in the case of *P(R|N&E) is low*

In any case, I am forging this theory of looming defeat entirely on my own. For all Plantinga says, *everything is a looming defeater for everything*, or else, he simply hasn't told us *what* is in need of deflection, and therefore, has not told us that believing premise 1 *requires* a deflector of the naturalist. What is clear is that, since defeater-deflectors do

not deflect *actual* defeaters, one cannot say that what is in need of deflection is *defeaters*. Defeaters are never deflected, because what is deflected, by virtue of being deflected, *was never a defeater in the first place*.

To return to Plantinga himself, Plantinga argues that the belief tha*t P(R|N&E) is low*, against a backdrop of belief in N&E, provides a defeater for R that itself has no defeaters, and a *looming* defeater for R that is undeflected, i.e, belief in the probability premise is an an undefeated, undeflected, defeater for R for naturalists. Plantinga claims that no defeater for R can itself be defeated, because any appeal to a putative defeater defeater -- for instance, a cognitive check-up at MIT -- would "presuppose that [the subject's] faculties are reliable" (2011a, 345).

This makes a certain kind of sense: once one finds oneself in a position in which one acknowledges that belief in R *is* irrational, then one cannot, while recognizing that belief in R is irrational, rationally work one's way out of that position. For in so doing, one would be employing a faculty that one, as we have already admitted in calling one's belief "defeated", cannot rationally believe to be reliable.

A *looming* defeater for R, however, may in principle be "deflected". And, since I have just now argued that everything is a looming defeater for everything, and every single belief is in need of deflection, I would *hope* deflection of at least some looming defeaters of R is possible. For if there could not be deflectors for any looming defeater for R, then *no one* could ever believe R rationally. And if the remaining premises of Plantinga's are true, then it would also be the case that *no one is ever rational in believing anything*.

Conditions on Defeater-Deflection

What are the conditions on the deflection of (looming) defeaters? Plantinga says (2002b, 224) that, for deflecting his putative defeater of R anyway, two conditions are obvious:

(1) The subject must hold the deflector belief, Q; and

(2) Q restores non-low probability to R, i.e., the probability of R, given the belief that *P(R|N&E) is low,* and also the probability of R given the belief that *N&E&Q,* is *not* low.

Beyond those two conditions, he says, "I certainly don't know how to give a complete and rigorous (or even a complete and unrigorous) answer to it" (2002b, 224).[30] Nevertheless, he does, in several of his presentations, offer further conditions:

(3) Q is not equivalent to R "in the broadly logical sense", e.g, Q is not the belief that (R or 1+1=1)) (2011a, 347)

(4) Q is not a conjunction of R & other of the subject's beliefs (2011a, 347).

(5) Q is not a proposition that no rational person in S's circumstances would believe (e.g., a contradiction) (2011b, 440)

(6) Q is not evidentially dependent upon R for the subject, i.e., it is not the case that the subject believes Q only on the evidential basis of R (2011a, 348)

Conditions (3) and (4) are designed to block question-begging deflectors. If (3) and (4) may be violated, then every looming defeater can, very easily, be deflected. (5) and (6)

[30] More recently, Plantinga hedges as follows: "This is not a trivial question, as one says when one doesn't really know the answer. But even if we can't easily come up with a rigorous statement of necessary and sufficient conditions for admissibility, we can still see some obvious necessary conditions" (2011b, 440).

only appear in one work each (and different works), both in 2011, so I am not sure which of (5) or (6) represents Plantinga's most recent view. The evidential dependence condition, condition (6), appears only in Plantinga's book, *Where the Conflict Really Lies*. The ban on universally irrational beliefs appears only in "Content and Natural Selection". The latter, the ban on universally irrational beliefs, can, I think, be safely added to the list without controversy: relying on irrational beliefs to save beliefs seems a sneaky way of avoiding any defeaters whatsoever, and Plantinga is right to close that sort of "loophole". The ban on anything one believes on the evidential basis of R, however, is anything but clear to me, as I shall explain shortly.

An "easy" response to Plantinga: candidate defeater deflectors.

With a common-sense reading of (6) it is obvious that (1)-(6) are easily met in the case of Plantinga's "looming" defeater for R for naturalists. Here are several deflectors, and I am sure the reader can supply many more along the same lines:

- Having true beliefs, rather than false ones, is adaptive;
- My friends have lots of true beliefs and they agree with me on most things;
- My *predictions* have turned out to be accurate;
- My *memory* is generally reliable;
- My *vision* is generally reliable;
- p_0, and I believe that p_0, p_1, and I believe that p_1, …; and
- There is *some* true naturalized theory of content, C, on which P(R|N&E&C) is high.

It should be obvious that all of my candidate deflectors meet conditions (1)-(5). (6) is the only possible sticking point. The phrase "on the evidential basis of" could mean many different things. On an *inclusive* reading, it could mean *presupposes*. On an *exclusive* reading, it could mean, *explicitly reasons through as a premise in causing or sustaining belief*.

Now, arguably, all of my candidate deflectors are believed on the evidential basis of R on the inclusive reading considered. For we *presuppose* R in everything we believe, at least if we are reflective believers. Every belief presupposes R, which is precisely the reason defeating R is supposed to defeat everything else we believe. And therefore, if there are any looming defeaters for R, and if these looming defeaters can only be deflected by beliefs that do *not* presuppose R, then there can be no such deflectors. Thus, naturalism really is defeated. *But so is everything else.* For, if as I argued earlier everything is a looming defeater of everything, then everything is a looming defeater of R. And no looming defeater of R can be deflected, in principle, given an inclusive reading of condition (6) on deflectors. So *everything* defeats R.

For instance, let us consider Plantinga himself. He believes not-N. In a noetic system with the belief that if not-N, then not-R, the belief that not-N would defeat the belief that R. So it is "looming" as a defeater for Plantinga. How does he avoid defeat? Well, he doesn't believe the conditional, *if not-N, then not-R*, of course! Right. He believes its negation, in fact: *it is* not *the case that if not-N, then not-R*. That belief deflects the looming defeater of R for Plantinga. Ah, but doesn't that candidate deflector get *disqualified* by (6) on the inclusive reading? For surely, every belief of Plantinga's

presupposes R, including all of Plantinga's candidate deflectors for the looming defeater, not-N.

Let us consider the other reading, the exclusivist reading, where evidential reliance is a matter of explicitly invoking something as a premise. Well, my reasons for thinking that truth is an adaptive quality of beliefs do not explicitly include R. I think knowing the location of water sources is adaptive, knowing the whereabouts of predators is adaptive, knowing how to find or build shelter is adaptive, and in general, for each kind of belief, I think the true beliefs are more helpful than the false ones. Where have I invoked R in this reasoning?

Much the same could be said of my belief that my friends are reliable, and that I share their beliefs. I do not invoke R (for myself) and then extend it to my friends, and then back to myself. Well, certainly not explicitly. I do, of course, presuppose my own reliability, but we have already set that observation aside as relevant only to the inclusive reading of (6).

Again, I do not think my memory is reliable because I am generally reliable and memory is one of my faculties. It may go in just the opposite direction. When I evaluate my general reliability, I do so by running through the reliability of each of my faculties. Memory? Decent. Vision? Quite good. Hearing? Excellent. And so on…

So, either way, the EAAN fails. On the inclusivist reading of (6), Plantinga's argument shows that no one has any rational beliefs. On the exclusivist reading of (6), the naturalist has plenty of deflectors available. Plantinga is thus right to say that his conditions are necessary, but *not sufficient* for disqualifying candidate deflectors. If they

were sufficient, the EAAN would have many, many easy responses. As we will see, premise 2, the defeat premise, is really not helped at all by this talk of defeaters, defeater defeaters, and defeater deflectors. For we have no theory of deflection that is of any real use to us in thinking through this particular case. Instead, everything rests on Plantinga's analogies.

Another possible restriction on deflection

We might consider other candidate restrictions on deflectors. For instance:

> (7) supposing S has a looming defeater d, threatening to defeat her belief in the reliability of a certain faculty, F, no deflector of d can itself be a product of faculty F.

(7) would disqualify my proposed deflectors, for what is "looming", namely, *P(R|N&E) is low* "threatens" R, and R includes all of my faculties. All of my beliefs are the product, therefore, of what has been threatened, and therefore beg the question against the defeater, and are thereby disqualified by (7) to deflect my defeater.

However, Plantinga never endorses (7), and he is wise not to. If everything is a looming defeater for everything, then everything threatens R for everyone. And (7) would disqualify any deflector for anyone. So, endorsing (7) is just endorsing radical skepticism for all. Plantinga leaves us, therefore, with no disqualifying criteria for all of my candidate deflectors. Still, he never claims that any deflector candidate that satisfies (1)-(6) is admissible. His criteria are *necessary* for genuine deflection but *not sufficient*.

In this situation -- with insufficient criteria for deflection that nab the naturalist, but not everyone -- it is clear that Plantinga's entire argument rests on the strength of his analogies.[31]

The admissibility of theories of belief, and their grounds

I will turn to the analogies shortly, but one more point must be mentioned before moving on from (looming) defeater deflection. Something else is curious in the discrepancies between the paper, "Content and Natural Selection" (2011b) and the book, *Where the Conflict Really Lies* (2011a). It is the book chapter that has condition (6), the puzzling "evidential basis" requirement. The paper, which omits that condition, allows that "Considered beliefs about the nature of belief itself can, presumably, be properly added, and RM is one of those" (440).

Throughout the paper, Plantinga considers various theories of belief on which P(R|N&E) is not low, and dismisses them one by one (2011b). He takes up, in turn, reductive materialism, functionalism, indicator semantics, and teleosemantics, dismissing each as potential defeater deflectors for one reason or another. But he dismisses them on the grounds either that the particular theory does not, in fact, raise the probability of R

---

[31] Omar Mirza is thus right when he says:

> The Defeater Thesis is at the heart of EAAN, and the most widely cited defense of this thesis involves an analogy with the XX case. But few philosophers have tried to challenge or investigate the intuition that is meant to be elicited by the XX case, namely the judgment that the subject in that case has a defeater for R. It is generally just granted, even if only for the sake of argument, that this intuition is sound. I conjecture that a deeper understanding of this case will lead to one of two results: either the intuition will be rejected, in which case the Defeater Thesis will be undermined; or else we will be able to determine which epistemic features of the XX case best explain the intuition. In the latter case we can investigate whether or not the EAAN case has the very same features, and hence whether or not it is genuinely analogous to the XX case (2011, 86).

(reductive materialism), or that it does not explain how N itself can believed (indicator semantics, teleosemantics) (Plantinga, 2011b). The complaint Plantinga does not even consider is that these theories violate (6) by somehow evidentially relying on R.

What is interesting is that if "considered beliefs about the nature of belief" are admissible, in the sense that they violate no stricture against "question begging", then it would seem all sorts of *other things* would be admissible in the same sense: they would pass condition (6), at least. For instance, let us ask what is the evidential support for theories of belief, generally? How do debates about the nature of belief go, and how are they settled?

Theorizing about belief

Having followed some of these debates over the naturalization of belief content, they seem to go pretty much like all philosophical debates. We have a vast store of pretheoretical beliefs about beliefs, ours and other creatures'. These beliefs are general and aphoristic -- *beliefs cause behavior, jointly with desire, what is known is believed, etc.* -- and they are specific -- *I believe right now that I have hands, I do not believe it is raining, a frog on a lily pad does not believe that either a fly is passing by or 2+2=4*. A theoretician proposes some simple principles, formed in a language free of terms for belief, and like notions. Those proposed principles and their consequences for particular cases are then critically compared to our pretheoretical beliefs about beliefs. We allow theories a certain degree of revisionism, provided they offer elegance, simplicity, and enough of a thrill of reductive explanation in return. But where the output of the

principles is very badly mismatched to our pretheoretical beliefs about belief, it is outlandish and we call that a "counterexample" to the theory. When we do this by making mutual adjustments to our general and to our particular beliefs, we call it *reflective equilibrium* (Goodman, 1983).

Like any sub-literature in philosophy, the theories naturalizing content are dazzling in their creativity and innovation, while the counterexamples to them are also clever and never-ending. The "naturalization of content", or as it is sometimes called, the "naturalization of intentionality", or "causal theories of content", is an enormous, and enormously difficult project. But these simple meta-observations about the nature of the field raise serious questions about Plantinga's idea of what is and is not admissible.

For, Plantinga admits that the theories of belief -- indicator semantics, functionalism, teleosemantics, etc. -- are admissible deflectors, in the sense that they do not violate any worries about circularity. (They fail to deflect for other reasons, *viz.*, they do not raise the probability of R or they do not allow that naturalism is a content of our beliefs.) But if belief in these *theories* does not in any way beg the question, then it would seem that all of the *beliefs we use to produce and evaluate theories of belief* would not beg the question either. But in that case, the vast stock of beliefs we have about belief, and which our theories are designed to capture, are admissible, at least in the sense that they do not violate any kind of non-circularity constraint. But then, that vast stock of belief about belief can, itself, be used to deflect Plantinga's looming defeater.

For instance, suppose a naturalized theory of belief content has, as a consequence, that right now I do not believe I have hands. Or suppose it maps a brain state I am in

right now to the belief that 2+3=27. I will reject that theory, of course. But then, if the theory is non-question-begging, wouldn't my belief that someone in my condition right now is, in fact, believing that they have hands, and is *not* believing that 2+3=27 also be non-question-begging? And I am assuming, of course, that naturalism is true, all the while as I make these judgments. So, why can't I use such judgments to support the belief that, *even given N&E, truth is adaptive*? N&E is true, after all, and I have so many instances where true belief is more adaptive than false belief, and I have the general belief that true belief is more adaptive than false belief. And even independent of my particular cases of adaptive true beliefs, I have the general belief that true beliefs are adaptive. And these beliefs are what motivate and what check my theory of content, for instance, it is what motivates and checks teleosemantic theories. Are such beliefs admissible? If so, the "looming defeater" for naturalism, namely, premise 1, is easily deflected. We do not need the *actual theories themselves*, all worked out. Their grounds will do.[32]

The other option is to say these grounds for theories of belief are *inadmissible*. But then, how are the theories admissible?

Plantinga, it is clear, wants us to screen off all of our beliefs about the actual world when considering premise 1, the probability premise. "We are asking about P(R|N&E), not about P(R|the way things actually are)" (2011a, 335-336) and he does not allow us any "centering" information in thinking about how things would be if N&E were

[32] Recent articles attacking the EAAN have tried to defend teleosemantics, for instance (Ye; Leahy). But teleosemantics faces a host of challenges. Must one really defend the theory in its specifics in order to believe anything rationally? That seems too tall an order, a point to which I will return in the final section of this paper.

true. Maybe he could leverage this same sort of requirement against deflectors, disallowing any candidate deflector from being a deflector of *P(R|N&E) is low* if it makes any assumptions that depend on *how things actually turn out*, rather than what is true, so to speak, regardless of how things actually turn out.

I think Plantinga might take this suggestion. Though my putative defeater deflectors do not violate Plantinga's explicit conditions on deflection, I think he would accuse such beliefs of being "cheating" and "inadmissible" for the same reason he accuses the belief that true belief is adaptive of being irrelevant to premise 1: it presumes things about the actual world, when we are judging the probability of R, given N&E, not given the way things actually are.

But at the same time, Plantinga allows that considered beliefs about belief itself are admissible, and I am not sure that general theories of belief could pass my proposed test (2011b, 440). The problem here is that our beliefs about the nature of belief at least appear to be very much shaped and supported or refuted by our beliefs about *how things actually are*. The actual world is full of supporting examples and counterexamples for theories of belief. So either one's views on the nature of belief itself must be *cleansed* of all support that could be *tainted* by the actual -- for instance, converting every counterexample involving the actual into one involving only a hypothetical, a hypothetical which itself is known a priori and independent of any knowledge of how things actually are -- or Plantinga should reject all theories of belief on anti-circularity grounds. Or, finally, he should admit the beliefs that support those theories of belief, in which case there are plenty of "deflectors".

These are not impossible choices. Plantinga could take my suggested restriction on deflectors for his defeater -- that they presume nothing about the actual -- and adopt a hyper-rationalist approach to theorizing. Maybe theories are supported in an entirely a priori way. But perhaps he does not want to shoulder such rationalist commitments. More problematically, his audience of naturalists will not: naturalists are rarely rationalists.

But I am two steps deep onto the path of "solving" Plantinga's problems for him, and perhaps Plantinga would take some other path. What I should say is simply that Plantinga's criteria for deflectors are easy enough to meet. There are many candidate deflectors for premise 1. And then I should put my point about theories and their supports as follows: Plantinga needs a sharper formulation of his anti-circularity requirement such that theories of belief *do not* violate it, but many of the judgments we use to support or "counterexample" such theories *do* violate it.

The analogical argument for the defeat premise

Plantinga never claims that his conditions for deflectors are jointly sufficient for deflection. Instead, his case for the defeat premise rests on an analogy, or rather, a family of such analogies. Since the theory of defeater-deflection is insufficient for Plantinga's argument -- and if taken to be complete, actually shows how the defeat premise fails -- and the analogies do all of the heavy lifting, I will set aside defeat talk for the remainder of the paper. Premise 2, in defeat-free language, simply says that it is irrational to believe R while believing *P(R|N&E) is low*, while also believing *N&E*. That is clear enough.

In favor of premise 2, Plantinga asks us to consider analogies. Because the entire argument hinges on these analogies, and because they are the locus of my criticism, I will cite Plantinga at length here. First, from *Where the Conflict Really Lies*:

> Suppose there is a drug -- call it XX -- that destroys cognitive reliability. I know that 95% of those who ingest XX become unreliable within two hours of ingesting it… Suppose further that I come to believe both that I've ingested XX a couple of hours ago and that P(R|I've ingested XX a couple of hours ago) is low; taken together, these two beliefs give me a defeater for my initial belief that my cognitive faculties are reliable. Furthermore, I can't appeal to any of my other beliefs to show or argue that my cognitive faculties are still reliable… Any such other belief B is a product of my cognitive faculties: but then in recognizing this and having a defeater for R, I also have a defeater for B. (2011, 342)

And in a footnote on the same page:

> Other analogies: the belief that I have mad cow disease and that the probability that my cognitive faculties are reliable, on that proposition is low. Similarly for the belief that I am a victim of a Cartesian evil demon who brings it about that most of my beliefs are false… and the current version of Descartes's fantasy, the belief that I am a brain in a vat, my beliefs being manipulated by unscrupulous alien scientists (see also the film *The Matrix*, Warner Bros., 1999).

Next, from a summary of the EAAN in a volume of criticisms and responses:

> 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 a 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 I have been 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. 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 witholding my natural belief that my cognitive faculties are in fact reliable (Plantinga, 2002a, 11).

The analogy goes as follows. Just as believing oneself to have taken the XX pill, or to be a victim of an evil genius, or to be in the matrix, or to be an experimental subject for Alpha-Centaurian superscientists, is a defeater for R, for oneself, so too, believing that *P(R|N&E) is low*, while believing *N&E*, is a defeater for R. Not only is premise 1 a defeater for R (for naturalists), but just like all of these other cases of defeat, it is a defeater that cannot itself be defeated by anything else, since it defeats in global fashion. Plantinga writes, about the XX case:

> Suppose, therefore, that I take a good dose of XX, which induces not merely perceptual unreliability but global cognitive unreliability. I believe that 95

> percent of those in this condition are no longer reliable; I also believe that 5 percent of the population has the blocking gene [and are unaffected by the drug]; but I have no belief as to whether I myself have that gene. I then have a defeater, so I say, for R. Now suppose I come to believe that my physician has telephoned me and told me that I am among the lucky 5 percent whose reliability is unimpaired by ingesting XX. Do I now have a defeater-defeater? Or do I still have a defeater for R? (2002b, 227)

Plantinga argues that in this case he still does have a defeater for R. He asks us to think of the case in the third person (2002b, 228). We are supposed to think about Sam, who we have never met. We assume R is true of Sam, but then learn he has ingested XX. Moreover, Sam believes he has gotten a call from his doctor to tell him he has the blocking gene. Should we still think Sam is reliable? Arguably not. His belief about the doctor's call is most likely the result of the drug, and not an actual call (2002b, 228).

Problems with the analogies

The cases Plantinga presents are, to my mind, very different from one another, and the differences between them matter a great deal to the argument. On some of the cases, I want to challenge Plantinga's judgment about whether there is indeed any defeat happening in the case itself. On others, I want to challenge Plantinga's claim that they are analogous to believing premise 1 to be true. Let me first set out the differences between the cases, as well as the ambiguity that infects some of them.

Distinguish: *certain* failure of reliability and *probable* failure of reliability.

Notice that some of the cases involve a belief that *all* the members of a certain class are unreliable, and that one is a member of that class, whereas other cases involve a belief that *most* of the members of a certain class are unreliable, and that one is a member of that class. In some of the analogous cases it is actually not clear which of these two possibilities we are supposed to be imagining. For example, the belief that one is being manipulated by an evil genius, if the case is filled in in the usual ways, is a belief that one is *actually being deceived*. That is, it is equivalent to a conjunction: *I am being manipulated by an evil genius in certain ways & not-R*. On the other hand, the belief that one has taken the XX pill, which only causes unreliability in 95% of its ingesters, does not entail not-R (for the ingestor).

This is an important difference. In cases where the scenario one believes to be actual *entails not-R*, it does not seem that one can go on believing R rationally while believing that scenario and seeing its entailment of not-R.[33] It's literally a case of believing *all Fs are G*, *I am an F*, and *I am not a G*. As any introductory logic student can tell you, that is very, very bad. So if we think about *those* cases, I agree with Plantinga's judgment about the case itself: those agents have defeaters for their beliefs. But about those same cases, I also think they are clearly disanalogous to premise 1. Plantinga has offered no reasons why N&E would *entail* not-R, nor has he claimed as

[33] One may have no choice but to go on believing R. Plantinga calls this "proper function rationality" and allows that you do not have a "proper function" defeater, because, hey, cognitive life must go on (2002b, 228). But one still has what he calls a "Hume defeater" (2002b, 228). There is still something *wrong* about what you are doing.

much in the argument. Therefore, cases of *certain* unreliability do not provide a basis for believing premise 2.

So we can ignore those cases and focus on cases in which one believes one is a member of a class *most* of whose members are globally unreliable. We can therefore ignore the demon case, the Alpha-Centaurian superscientist case, the Matrix case, and the mad cow case, if it is advanced-stage anyway, where unreliability is a sure thing.

There will be probabilistic versions of these cases we cannot ignore on these grounds, however. Maybe the genius loves lotteries, and randomly chooses 95% of believers to deceive, and *mutatis mutandis* for the other scenarios: the Alpha-Centaurian scientists love lotteries, mad cow disease spares some portion of its carriers, and so on. These probabilistic cases, for all I have said, may still work for Plantinga. But what Plantinga should *not* do is to leverage our firm intuitions that believing R is irrational if one also believes not-R (and some other things that explain not-R) to support premise 2. Of course one should not believe a contradiction. But the naturalist need not believe a *contradiction* to believe R. So let's narrow our focus to the probabilistic versions.

Clarify: what is the *phenomenology* of the subjects in the cases?

Because we are familiar with some of the scenarios Plantinga invokes, I think we tend to treat them all in the same way, namely, as scenarios in which the subjects have a phenomenology pretty much like ours: rich, coherent, consistent, and so on. And it is not just these general features of their phenomenology that we are imagining, in imagining the scenarios, but I think we imagine the scenarios in a way that the subjects are having

the very same phenomenological experiences as our own. So, when we imagine the genius hypothesis, we imagine it *as actual*, and *as actually true of us*.

But if we look carefully at the description of the cases, there is no mention of the coherence, consistency, or rich variety of the subjects' experiences. There is definitely no commitment to the idea that in these scenarios, the subjects may have *our* phenomenology. We are just told that they are globally unreliable. The cases themselves -- the content of the beliefs that one has about what scenarios one is in -- are silent on the nature of the *experiences* of the subjects in them. But settling the nature of the experiences in these subjects is crucial to the argument, as I will now argue.

First Horn: it is not part of the case that the phenomenology of the unreliable sub-population is much like my own.

Let's begin with mad cow disease. First of all, we are going to have to change the case a bit to avoid the problem of *certain unreliability* just discussed. Let us just stipulate -- this is philosophy after all! -- that a certain tiny percentage of those who have one of the human variants of mad cow disease, e.g., Creutzfeldt-Jakob disease, do not suffer any symptoms whatsoever. Now suppose that I ate a lot of BSE-tainted beef during a certain time period in a certain place. Suppose I have done a lot of reading not only on the internet but in medical libraries, and with the assistance of experts, about the disease and about the place and time where I consumed the beef and I have come, thereby, to believe that conditional on my diet, it is highly likely that I have the disease, and it is highly

likely that if I have the disease, at the advanced stage I would have it if I have it at all, I am *globally unreliable*.

Now supposing that, given the history of my diet, R is highly improbable for me, is my belief in R defeated? No! Not at all! Those who are in advanced stages of CJD, and whose cognitive faculties are totally corrupted by it, are indeed globally unreliable. But they have characteristic symptoms I do not have. They go globally unreliable *in a certain way*. I may be unreliable, even globally unreliable, but I know I am not unreliable *in that way*.

The Mayo Clinic lists the symptoms of CJD (Mayo):

- Personality changes
- Anxiety
- Depression
- Memory loss
- Impaired thinking
- Blurred vision or blindness
- Insomnia
- Difficulty speaking
- Difficulty swallowing
- Sudden, jerky movements

Suppose my personality hasn't changed, I have no anxiety, feel upbeat, see clearly, sleep soundly, can swallow with ease, and move as smoothly as ever. Well, suppose that is my impression anyway. Of course, it could be that my "impaired thinking" is leading me

astray on all of these counts. But that is not how CJD patients' thinking is, in fact, impaired. They suffer delusions, suppose, which have very similar, very characteristic forms. No CJD patients have *these* kinds of delusions -- delusions of sleeping well, swallowing easily, seeing clearly, moving smoothly, and so on. Given this setup, I would be quite confident, and rationally confident, that I am one of those who is unaffected by CJD, if I have it at all.

Let's consider another case, which should be clearer. Paranoid schizophrenics are globally unreliable. They are not only wrong in believing themselves to be the target of malicious parties, but they are wrong about their own reliability: they think everyone else is unreliable, but that they themselves are onto the truth: *conspiracy*! Now, suppose I take a pill that I believe induces severe paranoid schizophrenia in 99% of those who take it, within an hour, and which lasts for an hour. Suppose I am, and believe myself to be, 90 minutes post-ingestion, and I am feeling calm and untroubled. I don't think anyone is out to harm me, but find myself as trusting as ever. Should I believe that I actually have paranoid schizophrenia but have a false belief about having it? After all, it is 99% probable, relative to another one of my beliefs. And if I *did* have paranoid schizophrenia, I wouldn't believe I had it!

But I am not at all troubled, in such a case, and have no doubt whatsoever that I am in the 1%. We can make the odds even longer. I don't actually care *what* percentage of pill ingesters become paranoid; if I am delusional, I am not delusional in the *paranoid* way, because *I am not paranoid at all*. And I am not, right now, thinking that I am not paranoid in the way that paranoid schizophrenics think *they* are not paranoid. For there

are two ways not to be paranoid. The first is *not to fear at all*, and the second is *for one's fears to be actual*. I believe I am not paranoid in the first way, but those who suffer from the illness believe themselves to be not paranoid in the second way.

What is the *principle* here? What is allowing me to shrug off what looks like such a powerful "looming" defeater? The scenario I believe to be actual, and which makes me almost certainly unreliable, is a scenario that imputes to me a different phenomenology than the one I actually have, in cases where I am actually unreliable. Now, I might be a little bit off when making judgments about my own phenomenology. A skeptical scenario according to which I have mistaken an itch for a tickle cannot be ignored when I feel what I take to be a tickle. But, to borrow a case from Stephen Schiffer, if the scenario has the unreliable sub-population all seeming to themselves to be climbing a mountain right now, I simply don't have to worry about it (Schiffer, 336). It seems to me, right now, that I am writing a paper, nearly at sea level. And that seems nothing like climbing a mountain at all. In order to worry me, a scenario must describe a phenomenology not too different from my own. It must explain why I take myself to know things that I do not in fact know (Cross).

Bearing this in mind, return to Plantinga's scenarios. Notice that apart from the *mad cow* case, where I think I would not actually be worried at all, the other cases simply do not specify the phenomenology. The reader is left to implicitly or tacitly fill in questions about *what it would be like* if one were in one of the cases presented, and what it would be like in either in the reliable portion of the population or in the unreliable portion of the population in the case one believes one to be in.

I have just made an argument that if the analogies are supposed to *work*, that is, if they are supposed to generate the intuition that the subjects who believe themselves to be in the scenarios are *irrational* if they continue to believe R, of themselves, then the unreliable population in the scenario should have a phenomenology close to the one the subject takes herself to have. Otherwise, the subject can comfortably and rationally conclude that she is in the reliable sub-population within the scenario.

Let's consider the XX case. We are only told that 95% of ingesters are unreliable within two hours of taking it. We are not told in what *way* they are unreliable, at least in this presentation (Plantinga, 2011a). So here, one possibility is that the drug puts 95% of its ingesters into a coma, where they form only one belief: I am itchy! If the drug works like that, and I am not at all itchy, nor tempted to believe I am itchy, I can safely assume I am in the 5% of non-responders. Of course, another possibility is that it works the way the evil genius does, by providing to the ingester of the XX pill a whole simulated "world" that corresponds to nothing outside of the mind of the ingester. If the pill works in that way, I am indeed worried that I am in the 95%, if I believe I have taken it. That sub-scenario *explains* why I, right now, take myself to be a non-responder when I am in fact a responder. The first sub-scenario *did not explain* why I take myself to be a non-responder when in fact I do respond to the XX drug.

How should I respond to the case, not knowing anything more than that the drug makes 95% of ingesters globally reliable, i.e., not knowing *how* it makes them unreliable? Well, my instinct is to assign probabilities to the sub-possibilities, first, that I have taken a drug that erodes reliability in a way that is phenomenologically identical, or

nearly so, to my current experiences, and second, to the sub-possibility in which the pill works by causing a radically different phenomenology. I will then worry myself only to the extent that I assign a high probability to the first sub-possibility.

As a matter of fact, if I think about having taken such a pill, I would not be worried at all. We are so very far, in pharmacological technology, from developing drugs that cause *this* kind of hallucination, hallucinations or delusions with *this* very phenomenological content -- the kind I am having right now -- that such a drug would be a *spectacular* scientific discovery. I would surely have read about it in the news. By contrast, causing 95% of ingesters to be globally unreliable is actually *not very hard*, provided one gives them a phenomenology nothing like mine. Lots of drugs already induce delusions, and those delusions can be severe and pervasive. Thus, XX is almost certainly one of those familiar delusion-inducing drugs with a longstanding entry in the pharmacopeia. I may even know what those drugs are like through previous firsthand experience, and I may know that I am not having the kinds of experiences people have on one of those drugs.[34]

Second Horn: the phenomenology of the unreliable sub-population is indistinguishable from my own

Let's fix the cases to avoid my concern. Let's specify, in the cases, that the phenomenology of the unreliable subjects *matches* the phenomenology of the reliable

[34] I see nothing epistemically wrong with such reasoning, and I suspect most ordinary non-philosophers and most philosophers too, would agree with me. Perhaps a confirmed anti-internalist like Plantinga might have a problem, but rhetorically, he will have an uphill battle if his EAAN depends on his anti-internalism.

subjects. So, let's just stipulate that you believe you have taken XX and you believe XX works by creating experiences exactly or almost exactly like yours, and also making 95% of those experiencers unreliable. And let's stipulate, similarly, about the probabilistic version of the Alpha-Centaurian superscientists and the evil genius, and so on, that they make a complex, rich, consistent, set of experiences that are indistinguishable from the ones had by the reliable minority in each case. So specified, my intuitions align with Plantinga's on the cases themselves. In each case, it would seem, the subjects who continue to believe R are doing *something* wrong.

However, now a new problem arises for these cases: they are no longer clearly analogous to premise 1. They may still be analogous, but it is not obvious, and that is enough to disrupt the EAAN. Think about the way the argument went for premise 1. Assume N&E. There are subcases of the scenario under consideration (N&E) where believers are reliable and subcases where believers are unreliable. We are asked to conclude that there are more subcases where believers are unreliable than subcases where believers are reliable. And we are supposed to think this on the grounds that all different assignments of belief contents to neural states are equiprobable, and thus, we should give equal probability, for a random *adaptive* neural state, to a content assignment that assigns a true belief to that state as a content assignment that assigns a false belief to that state. The probability that most of one's beliefs, say 90% of one's beliefs, get mapped to true propositions, is low, given that each belief has roughly a 50% chance of being mapped to a true proposition, and that one has many, many beliefs.

Notice that nothing in that argument tells us explicitly *what it is like* to be one of the unreliable believers, or, for that matter, one of the reliable believers. We are thus in something like the situation we were in when considering the XX pill case, which was silent on the phenomenology of the ingesters. If one believes that the likely ways in which R is false for a sub-population are ways one can phenomenologically distinguish from one's own experience, then, just as we located ourselves in the 5% of non-responders for the pill, we could locate ourselves, unproblematically, in the minority of believers for whom R is true. And one very well might believe that if R were false, things would probably seem very different.

Here is a strategy a naturalist might employ. Let's consider those arbitrary semantic assignments. How would things seem if belief contents were assigned willy nilly? That is, how would it seem if the brain state I am in when I am faced with a tiger were the belief that 2+2=4, or that the EU will inevitably dissolve? Or how would it seem, even, if I were believing that tigers are unlikely to eat me, while at the same time desiring to be eaten?

One answer is that things would seem exactly as they in fact do; seemings would not differ at all. But that is hard for me to believe. Seemings are, themselves, belief-like states, if not actually kinds of beliefs: both seemings and beliefs represent the world as being a certain way. I would think a good theory of content would, ceteris paribus, coordinate assignments of *its seeming that p* to one and one's *believing that p*. The coordination will *not* be a perfect match, because there are cases where one believes contrary to the way things seem, i.e., when one resists the temptation to believe what

seems to be the case. But when that happens, the disconnect between believing and seeming *itself* feels a certain way, and that is certainly not the norm. Believing what seems to be the case feels different from resisting belief in what seems to be the case, because of overriding reasons, e.g. resisting the impulse to think that a bent-looking stick in water is actually bent, or resisting the impulse to fall for various other illusions, whether sensory or cognitive.

So, if beliefs and seemings are somehow coordinated, I think it is at least reasonable to judge that if arbitrary semantic assignments were *in place*, so to speak, things would *seem* very different than they in fact do, for the seemings would track the beliefs in their arbitrariness (again, ceteris paribus). I do not think one could simply leave the way things seem entirely fixed to the way they actually seem right now, but assign radically different belief contents, e.g., one could not assign my neural states right now to the belief that I am climbing a mountain without changing *something*, locally or globally, about the way things seem to me. (Or, at least, only a defective naturalized semantics would so radically disconnect seemings and beliefs.)

Recap of the dilemma for Plantinga's "analogous" cases

So, let me summarize my critique of the analogies thus far. First, the analogies must be put in probabilistic, not absolute form. Cases where subjects believe something that entails not-R, and also see that entailment, and yet who believe, regardless, that R is true, are irrational. But that tells us nothing about the status of naturalists who believe premise 1 and still believe R, because naturalism does not entail not-R.

Second, the analogies must be restricted to cases in which one believes oneself to be in a class of believers most of whom are unreliable, but also a class whose unreliable believers are phenomenologically very close, if not identical, to oneself. Otherwise, while the analogy between naturalists who believe premise 1 and the subjects in these "bad" hypothetical cases may hold, the subjects in the cases can rationally believe R, because they can rationally believe themselves not to be in the "bad" subpopulation in the case.

Third, once the cases are so restricted, it is no longer obvious that they are *analogous* to the case of a naturalist who believes premise 1. For, it is not clear, in the argument for premise 1, that believers who are unreliable are phenomenologically indistinguishable from ourselves. We know only that Plantinga considers arbitrary assignments of belief content to be equiprobable, but have been told nothing about what it is like for, or how things seem to, subjects if there are bizarre and unintuitive assignments of belief contents. Given very plausible constraints on aligning assignments of seemings with assignments of beliefs, it is also plausible that radically different assignments of neural states to belief contents from those we take to be actual *would* also *seem* very different. Since the naturalist has no reason to think things would seem the way they actually seem if assignments were radically different from the way we think they are -- a way in which true belief *is* adaptive and a way in which R *is* true -- the naturalist may continue to believe R, despite believing that R is improbable, given only N&E.[35]

[35] One can imagine scenarios that are phenomenologically alike, but where the contents of beliefs, or even seemings, differ, and where there are widespread errors as a result. Ned Block imagines

Recall that Plantinga's entire argument for premise 2, the defeat premise, is an appeal to these analogous cases. His explicit criteria for defeater-deflection, as I showed with examples, actually allow lots of deflectors for his "looming" defeater of premise 1. Everything rides on the analogies. But now, the analogous cases have proven ambiguous in at least two ways. Once disambiguated, either the cases fail outright or they fail to be analogous. We can conclude that Plantinga has not yet made a case for his conclusion: the EAAN has failed.

A return to the two premises and two kinds of critique

The probability premise really says that nothing in the *nature of belief itself* secures our reliability, given only that we have been selected by Darwinian struggle and that we are material beings. To dispute that premise, one must argue that something about the nature of belief really *does* make our reliability probable under conditions of Darwinian struggle

---

such cases in "Inverted Earth" (1990). A subject is, unbeknownst to himself, sedated, and has surgically inserted into his visual system a spectrum inverter. He is, at the same time and also without his knowledge, whisked away to a sister planet of Earth, where all of the actual colors of things are inverted: ripe tomatoes are green, the sky on a clear day is yellow, bananas are blue, and so on. When he awakes, he notices no difference from the previous day, but in fact, all of his color beliefs are now mistaken.

More generally, one can imagine inducing error in subjects without changing phenomenology simply by shifting the "wide" content of belief assignments. Plantinga could avoid my criticism by, for instance, specifying that our beliefs about water may in fact be about XYZ, and the residents of Twin Earth may have water beliefs about H20, so that neither of us has beliefs about what is, in fact, in our environments. So, of course there are still assignments of content that preserve phenomenology, yet have us going very, very wrong.

But now, it just does not seem equiprobable that what I take to be my water beliefs, for instance, are about H2O and that they are about a substance only found on Twin Earth, but not anywhere on Earth. And it does not seem probable that I have been operated on recently by scientists, unbeknownst to myself, and had a spectrum-inverting implant, and that again, without my knowledge, I have been whisked to a planet that is color inverted relative to my home planet. And these scenarios do not seem probable, even when I try to screen off what I know of the actual and just imagine what is likely given N&E.

and materiality. The target of dispute here is beliefs generally. Is belief *generally* true, in a world without design? An immense question, very difficult to answer, because we are acquainted with such a small slice of possible material, evolved, belief.

The defeat premise says that if we answer that broad question about belief in general, for all possible evolved creatures, by saying that no, reliability is unlikely *in general*, then we must subsume our own reliability under the generalization: we must assume for ourselves the same reliability as *all possible evolved believers*. We are not allowed to carve out for ourselves an exception space, in which we are more reliable than a randomly chosen, evolved, material, believer.

The two most prominent ways of attacking Plantinga's argument are to attack these two premises. The first kind of attack argues that beliefs, in general, really are more likely to be true than false under material, evolved, conditions. Stephen Law's criticism, that there are conceptual constraints tying the adaptivity of beliefs to their truth, is one such attack, but there are others. In "Content and Natural Selection", as I have said, Plantinga canvasses and rejects various naturalized theories of content: indicator semantics, functionalism, and teleosemantics (2011b). Plantinga treats these as "defeater deflectors", and thereby, as objections to his defeat premise (2011b). But really, they are theories that have as a consequence that belief, in general, under evolved, material conditions, is probably not mostly false. They are still about belief in general, and not just human beliefs in a world like our own.

The second kind of attack grants for the sake of argument that nothing about the nature of belief *itself* secures the reliability of believers, but denies that *we* should adopt

the same judgment about *ourselves*, as we now are, with our current experiences, as we do about some hypothetical believing creatures, whose inner lives, whose worlds, may bear no resemblance to our own. This second kind of attack is cast as an attack on premise 2, by way of providing defeater-deflectors in the form, say, of our beliefs about our experiences. Then the question is whether such beliefs are admissible deflectors, by Plantinga's rules, and since his rules are incomplete, the question becomes what is admissible in the analogous cases.

My own criticism of the analogies is of the second type: the analogies are invoked in support of the defeat premise, after all, not the probability premise. I am not arguing that belief, generally speaking, is reliable, but only that I have no reason to take *myself* to be unreliable, even if belief in general is likely unreliable for all possible evolved creatures, unless I *also* think that the subset of believers with the sorts of inner lives that I am now having are mostly unreliable. And further, nothing in the EAAN has supported that idea, only the very broad *general* idea that possible, evolved, material believers are mostly unreliable.

We have these two styles of critique, but it is still somewhat difficult to pin these two critiques directly on the two premises. For, Plantinga and Miller seem to pin both styles on premise 2. In fact, how we should do the bookkeeping depends on what we understand to be included in premise 1. Let's get more precise about the premises.

Ambiguities in premise 1: *P(R|N&E) is low*

Recall that N is highly indeterminate, as already discussed, because it says there is "nothing like" God. But that indeterminacy does not undermine the argument, so far as I can tell. What matters is that N entails or includes materialism, and I am happy to allow there is a kind of naturalism that *does* include materialism.

Notice that R is used in a variable fashion. In the argument it is about "us" and our reliability, but it is often used by Plantinga -- and in the present paper, by me -- to refer to specific individuals: *my* reliability, the reliability someone who believes N&E believes themselves to have, etc. And this step to the individual is important. For I may think that humans, generally, are unreliable, but that *I* am special. (Note: I see no problem in relativizing R to different populations, so long as it is always clear which population we are talking about.)

Notice that R cannot be construed as a sort of statistical measure over the collective output of the faculties, or the argument will not go through. For, suppose that I have one or two reliable sub-faculties, but mostly unreliable faculties. Then it may be the case that I am generally unreliable, but I could still limit myself, in principle, to the reliable faculties in supporting N&E. Not-R should thus be construed strongly as the thesis that *none* of our sub-faculties are reliable, and consequently, R should be construed weakly as the thesis that *some* of our sub-faculties are reliable. The argument for premise 1, if it works, would give all sub-faculties equal probability of unreliability. Again, this ambiguity poses no in-principle difficulties for the EAAN.

Ambiguities in E

E is ambiguous in multiple ways. Plantinga says E is "... the view that our cognitive faculties have come to be by way of the processes to which contemporary evolutionary theory direct our attention" (2011b, 435) and that E is "...the proposition that we and our cognitive faculties have come to be in the way proposed by the contemporary scientific theory of evolution" (2011a, 317). Where is the ambiguity?

Well, the first ambiguity is in *what gets included in contemporary evolutionary theory*. A wing of evolutionary theory is actually about belief itself: the origins of beliefs, their utility, their accuracy or inaccuracy in various contexts, and so on. Plantinga might argue that when evolutionists study, for instance, communication in packs of dogs, or cognition in crows, or cooperation in schools of fish, they are merely studying *indicators* and not beliefs. But this would be very hard to maintain in the study of deception and self-deception. Robert Trivers, one of the most prominent evolutionary social theorists of his generation, has devoted much of his career to the these subjects, and his explanations of what self deception is, and how it evolved -- in brief, self-deception enables deceit without the characteristic markers of deceit, which enables unpunished, non-cooperative behavior in a cooperative species, which, in turn, is adaptive for the non-cooperative individual -- are part of evolutionary theory, broadly construed (Trivers, 2011).

In fact, Plantinga has been criticized on the grounds that evolutionary theory, but not theism, best explains our particular mixture of reliability and unreliability (Childers, 2011). We are reliable in concrete matters relevant to our survival and less reliable on highly abstract matters, and that is to be expected on a godless materialistic story of our

origins, but not on theism (Childers, 2011). If we include such evolutionary stories about belief itself -- including belief's relative accuracy on different subject matters, how much self-deception is still adaptive, which inaccuracies are maladaptive and which are adaptive -- in E itself, then P(R|N&E) is not low after all. For E already takes a stand on R: R is already cooked into what we are conditionalizing upon.

What Plantinga needs for the EAAN is something like contemporary evolutionary theory scrubbed of any belief talk: E, minus anything intentional. Let's call it "E-". If evolutionists were to join the program of naturalizing content, and restrict themselves to a vocabulary that passes muster in a naturalistic analysis of content, then they, arguably, could *not* theorize about deception or self-deception, or if they did, it would to be couched entirely in terms of a naturalized analysis of such talk. Now, Plantinga can still run the argument with E- substituted for E throughout. And notice that anyone who believes E, believes E-, for it logically weaker.

Now, relative to our new argument with E- substituted for E, appeals to evolutionary theory insofar as it explains the origins of belief and deception and self-deception, are actually defeater deflectors, and relevant only to premise 2. Trivers is not doing evolutionary theory, *per se*. Or, I should say that what he is doing would be, from Plantinga's point of view, offering deflectors for a looming defeater that is *P(R|N&E-) is low*.

The danger of including too little in E

There is nothing wrong with substituting E- for E, though it does call attention to the slipperiness of this form of argument. For one wonders why one cannot weaken E- further still. E- contains the information that we exist. We have "come to be" by way of evolution. But why are we allowed to conditionalize on this? Evolutionary theory might be thought of us as a theory of *our* history, and the history of *existing* organisms generally. Hence, biologists study the details of *actual* species and *actual* lineages. The theory of evolution is a theory about how life, as it is, came to be. But also, evolutionary theory is general. When we run computer simulations, we are not necessarily trying to learn anything about existing organisms, but about *rules and principles* about reproducing populations in general, actual or not. We are trying to learn about how matter randomly comes into complex arrangements and entropy decreases locally even as it increases globally. What are the principles, or laws, of systems with random variation and natural selection? Let's call those questions about *general evolution*, as opposed to *specific evolution*. Another way of weakening E, then, would be to make it about *general*, rather than *specific* evolution: contemporary *general* evolutionary theory is true. And indeed, the word "evolution" can abbreviate both. Call this "E--".

Now, what is P(R|N&E--)? I would think it is *exceedingly* low, for it is highly improbable that "we" would have come be at all, given only the general principles of evolutionary theory. If we do not exist, we are not reliable and R is false. So if we ran the argument with E-- substituted for E in premise 1, premise 1 would clearly be true.

We should go just a bit slower. The sub-argument for premise 1 is as follows: the probability that I exist, given N&E-- is very low. Therefore, I have a defeater for my

belief that I exist. If I have a defeater for the belief that I exist, then I have a defeater for R, which presupposes my existence. But then if I have a defeater for R, I have a defeater for anything, just as Plantinga himself claims.

Does *P(I exist|N&E--) is low* really defeat my belief that I exist? It is a looming defeater. And what can deflect it? Here, I can't speak for Plantinga, since he gives us no sufficient set of rules for admissible deflection. What I can say is that anything to which we might appeal to deflect it seems to presuppose that I exist -- for instance, that I am thinking, or that I am walking, or whatever -- and thereby "beg the question". But if what qualifies, say, noticing that I am thinking, and therefore, that I exist, as an admissible deflector, is its clarity and distinctness, or its indubitability, or its having Moorean status, or any other such marker, then one might also appeal to the sorts of deflector candidates for the EAAN I offered earlier -- general and specific beliefs about beliefs, but also beliefs about how things seem to me now, and the coordination of believings and seemings -- and support their admissibility on the same grounds. And one might appeal to these even though they are *contingent*, and do not concern only a priori knowledge of the nature of belief itself.[36]

---

[36] The problem applies to theists too, not just naturalists. Suppose you think you were created by a voluntarist god. God was free to create you or not to create you. And the world would be no better and no worse if God made someone else instead of you. In fact, God chose from an infinity of possible creatures only slightly different from you, or maybe even qualitatively just like, but not identical to you. Your odds of making the creation cut were 1/infinity, let's say. But you believe that God did create you, of course, of her own free will, and created none of your alternatives.

So here is a probability claim that you should accept: the objective probability that you exist, given that God created the world is arbitrarily low. If you don't exist, you don't have faculties, and therefore you don't have reliable faculties. *So P(I exist|God created me from a wide range of alternatives) is low*. But so what? Obviously, this is not a defeater for all of your beliefs. Why not? Well, you have a deflector, e.g., that *you think*.

Similarly, given the difficulties of explaining consciousness from a material base, one might think that the probability of us being conscious, given N&E-, is very low. Then we have another looming defeater, now of our belief that *we are conscious*. For we believe things about our origins on which our being conscious is highly improbable. But what can deflect that defeater without begging the question against it? Won't any candidate deflector, for instance, that I am now conscious of the tapping of woodpeckers in the canyon outside my office, *presuppose* that I am conscious?

Reflecting on these ways of weakening E is important for two reasons. First, it shows that Plantinga will need to admit deflectors that seem to beg the question against the looming defeater. Otherwise, our beliefs that we exist and that we are conscious will be defeated by versions of premise 1 that weaken E. Secondly, it shows that premise 1, given the right weakening of "E" is surely true, but that premise 2, for some of these "Es", is surely false.

Here is the functional relation between the strength of E and the plausibility of the two premises: the plausibility of premise 1 is inversely proportional to the strength of E; the plausibility of premise 2 is directly proportional to the strength of E. As we strengthen E, the probability premise becomes less and less obvious, but the defeat premise becomes more and more plausible. As we weaken E, the probability premise becomes more and more obvious and the defeat premise becomes less and less plausible. [37] Plantinga needs a middle ground that provides a looming defeater for the naturalist,

[37] Otte does not put matters exactly like this, and does not target E in particular but I think he sees the same tradeoff (1992, 142). Otte also takes experiences to be what we should conditionalize on, though he does not think that hurts Plantinga's case in any way (1992).

but not for himself, and which cannot be deflected. It is unclear whether Plantinga has indeed found such a middle ground. But once we see the functional relation, we see that each premise is actually highly plausible, and each is highly objectionable, given a certain understanding of the setup.

What about my beliefs about E itself?

So far, we have settled on E- as the right way to understand Plantinga's "E". E- entails not only general truths about evolutionary mechanisms, i.e., E--, but also information about *our existence* and about *our evolutionary history*, but all references to the intentional have been expunged from that historical theory. Here, it is still tricky to say is meant by "us", or what we are supposed to conjure when thinking of "us".

Suppose what Plantinga signifies with "us" is the plurality of actually existing humans. Then it would seem that at least some of my beliefs are true, given E. For, I believe that Mark exists, and Meg, and Paul, and Steve, and the other Mark. And E includes that all of us who actually exist, exist, then given E, all of those folks exist. So given E, my belief that they exist is true. Also true, on E, are my beliefs that humans exist, that they evolved, and so on. In fact, if I am knowledgeable about human evolution, then on E, each one of those evolutionary beliefs of mine is true. And many more beliefs of mine may also turn out to be highly probable, on E, if I have lots of true beliefs about *other* species from which we descended. For then E will contain lots of facts about those ancestors as well.

Since I believe E, I think that, given E & anything else, however my reliability may be in general, at least my beliefs about E are true. So, even in the setup, I am not just in the great class of unreliable believers produced by the general processes of evolution (E--), for they may or may not have true beliefs about their origins, whereas part of what I am conditionalizing on is that E is true. I am *right* that human beings exist, that I am one of them, that all of my acquaintances exist, that we descended from earlier primates, and before that, other mammals, and before that, fish, and on and on until the very first living things, and presumably, before that to the conditions under which life arose, and the physics that allowed that to take place and the astronomical facts, and on and on. That is actually a lot of true beliefs! How strange would it be that I am so right about all of that, and yet in general, none of my faculties are reliable at all, given N&E? Wouldn't a simpler and more explanatory hypothesis be that I am reliable about a wide variety of subjects, in addition to N&E?

Now, I imagine that Plantinga would say the problem is that I am assuming that I believe E, and I am not entitled to that assumption in my probability judgment, *given only N&E*. For, if N is true, then materialism is true. And if we are material beings, then beliefs are just neural states and the adaptivity of those neural states is independent of their truth (not in terms of metaphysical possibility, but in terms of probability). But, and here is the point, as the target of the EAAN I think N&E is actually true, and I take myself to believe E. Therefore, if I consider the possibility that N&E is true, I simply think of the actual world, and in the actual world, I believe E, and E, and by extension, those beliefs that constitute E are true, of course, conditional on E.

Here, once again, we are up against the question of how to think of the objective probability of *our* being reliable, given N&E. Plantinga's procedure is to think about *some hypothetical creatures* who believe something or other, but not necessarily what we believe, and then to estimate the probability of R for ourselves as the *same*. But if we are conditionalizing on E, it is not clear this is the right procedure for estimating *our own* reliability. For we know something about ourselves, to begin with, that we do not know about them: we believe E! Why should we ignore that information about ourselves in estimating our reliability, *given E*? (And in fact, don't I need to recognize that I believe E in order for the argument to work against me, and convince me to change my belief?)

But perhaps we are reading too much into "us" and "me". Let's grant Plantinga that all the facts about yourself beyond your existence, your being human, your having beliefs, and your having originated in accordance with E-, are off limits for the purposes of evaluating P(R|N&E-), because "us" carries none of that information. Then suppose Plantinga is right that P(R|N&E-) is low. But now, the defeat premise looks weak. For why should *I* be moved by this general fact about what is probable, when I have information missing from N&E- alone, namely information about what I believe, including that I believe that E-? And since I know I believe E-, I know I believe a great many truths, conditional on E-. And therefore, I know something that sets me apart from the run-of-the-mill product of N&E.

If on the other hand "we" indicates us-with-the-inner-lives-we-already-have, then the probability premise looks false to me. For, with this inner life, believing in E- as I do,

I have a lot of true beliefs, and consequently, probably some reliable sub-faculties, and therefore R does not look improbable on N&E- after all.

Summary of these objections

Once again, through slightly different channels, we have reached the same dilemma for Plantinga: weaken E and weaken premise 2; strengthen E and weaken premise 1. The objection to the EAAN is fundamentally the same, but finds its way into premise 1 or premise 2, depending on whether we are really thinking of the probability of R for us, as we are, with the inner lives and beliefs that we now have, or whether we are thinking of the probability of R for us, while disregarding any of the things we know about our inner lives. The objection is that given the way our inner lives are -- our having these very experiences, our believing N&E, etc. -- P(R|N&E) is not low. Here, it is not because we would have different experiences if we were in the unreliable sub-population, but because we might have different beliefs. We believe N&E. The unreliable sub-population may or may not. And we are conditionalizing on N&E, so we know that those beliefs of ours are in fact true. Therefore, the unreliability of the sub-population that may or may not believe N&E does not clearly bear on our own, for we clearly do believe N&E, and therefore we clearly do have a great many true beliefs.

Plantinga will treat our beliefs about our inner lives, whether it is our beliefs or experiences or seemings, as candidate defeater-deflectors, no doubt, and dismiss them as inadmissible, because they in some way *beg the question* against the looming defeater of *P(R|N&E) is low*. I have argued that these beliefs meet all of Plantinga's explicit criteria

for admissible deflectors -- the union of each set he requires of deflectors in some publication on the topic -- under the only non-apocalyptic disambiguation of requirement (6). The only thing that stands in the way of these deflectors, possibly, is the *analogous cases* -- the XX case, the evil demon case, and so on. Everything rests on those cases. But, as I argued, they are ambiguous. On one disambiguation, they simply fail to generate the intuition they are supposed to generate: that the subject in the case cannot rationally believe R of themselves. On another disambiguation, they succeed in generating the intuition that the subjects cannot rationally believe R of themselves, but they are no longer clear analogues of us when we believe premise 1. For it is part of the setup that those subjects believe they are in scenarios in which they are probably unreliable and in which, if they are unreliable, they are in a state phenomenologically indistinguishable from their actual state. It is not obvious, however, that to believe *P(R|N&E) is low* one must believe that the possible unreliable products of N&E have inner lives matching my own. They may, most of them, have inner lives that are radically different from mine. Absent that stipulation, however, I do not see why P(R|N&E) is low should dictate the probability of R *for me as I know myself to be*.

Note that any argument that P(R|N&E) should so dictate the probability of R for *me*, simpliciter, must either: (a) embrace the equivalent response to the analogous cases in which we think it is unlikely that things would seem *this* way if I were in the unreliable sub-population in the scenario; or (b) argue that radically different content assignments *would make no difference to phenomenology*. I am not optimistic about either strategy,

though I am not *certain* that (b) won't work either. A fan of the EAAN could presumably take up the argument precisely at that point, though it has yet to be done.

Thus ends my critique. But I made two more promises at the opening of the paper: first, to explain an unappreciated fact about the strange dialectical situation for a naturalist who thinks the EAAN is sound; and second to show how the EAAN turns on four broader philosophical questions.

The fate of a naturalist who believes the EAAN to be sound

The EAAN provides no rational way out for the naturalist who acknowledges its soundness, not even the kind of prudential way out that Pascal offers his readers in his "wager" (Pascal).

Consider first that for all the EAAN says, it may turn out that naturalism and evolution are both independently *highly* probable and each makes the other overwhelmingly *likely*. If sound, the EAAN would remain sound even if *all* of the available empirical evidence – the fossil record, the evidence of gratuitous evils, and so on -- points towards naturalism and evolution. Indeed, so far as I can tell, the argument would remain sound even if there were to be, somehow, *a priori* proofs of both theses. It would still show that evolution and naturalism is, for a sufficiently reflective person anyway, epistemically forbidden fruit. And that is because the EAAN operates on a channel independent of the first-order evidence for or against evolution and naturalism. The probability premise would be true regardless. The defeat premise would be true

regardless. The other premise, spreading defeat to all of one's beliefs, including N&E, would be true regardless.

Let's suppose, then, that our naturalist has judged that the evidence, in fact, favors naturalism and evolution. Suppose she happens upon one of Plantinga's books, reads the EAAN and is utterly persuaded by it, and therefore decides to give up on all of her beliefs, including E&N. What is she supposed to do *next*? Rationally, she can do nothing, since she has a defeater for all of her beliefs, including her beliefs about how to get her doxastic house in order. She certainly cannot adopt *theism*.

Consider an analogy. Suppose there is an XX-pill-lottery. A billion people play the lottery. Nothing happens to the winner. The losers are all given the XX pill. This version of XX fills each and every ingestor of the pill with delusions which are indistinguishable from veridical experiences, and very much like the experiences had prior to taking the pill. The pill also erases memories of its ingestion, so after one loses the lottery and takes XX, one has no recollection of having lost the lottery. Now, suppose you play this lottery, remember or at least seem to remember playing the lottery, and you wake up the next morning and ask yourself whether you won or lost. Well, you almost certainly *lost* the lottery and have taken the pill that produces mostly false beliefs. The chance is 999,999,999/1 billion. But if you believe that you lost, and thereby that you took the pill, then you also have to believe that your faculties are *right now producing mostly false beliefs*, for that is how you believe XX to work. If your faculties are right now producing mostly false beliefs, then you should not believe that you lost the lottery.

But here is *also* what you should not do: you should not believe you actually *won* the lottery! That is a 1-in-a-billion shot. And your only reason for believing you have won is that losing would be very bad for your epistemic life. That depressing fact, however, in no way increases the probability that you won. The odds of your having won haven't budged; they remain precisely 1 in a billion. Noticing that your life would be better if you were the winner and believing that you won on that basis would be a textbook cases of wishful thinking: baseless and irrational.

Likewise for the EAAN. Anyone who believes the negation of naturalism on the basis of the EAAN is behaving epistemically irrationally. For they have changed their credence in naturalism solely on the basis of wishful thinking, hoping to hear good epistemic news about themselves.

So, what is a naturalist convinced by the EAAN to do? She plays backgammon, or sleeps, and hopes to forget about the EAAN. But what happens when she returns to consider E&N once more? She falls into the same trap again, assuming her beliefs in N&E have returned. (And why wouldn't they? The evidence points in that direction for her and she remembers her previous beliefs.)

Could she set up a hypnosis program, when she is freshly defeated, so that when she returns from treatment she will no longer believe E&N? Not rationally, no. If she recalls the EAAN, all of her beliefs are defeated, including beliefs about how hypnosis causes beliefs, how without E&N in the way, she could rationally believe other things, and so on.

Could she set up a hypnosis program at some other time, say, a while after her universal defeat? Not in rational response to the EAAN, no. For, if she does not remember the EAAN well enough to be defeated by it, she does not remember the EAAN well enough, rationally, to take steps to avoid defeat by it. But if she does remember it well enough to take (rational) steps to avoid defeat by it in the future, then she will be defeated by it in the memory of it.

One might try to imagine a kind of middle ground where the subject has just enough of a recollection of her self-defeat to want to avoid it, but not a sharp enough recollection to actually be defeated. But if the argument is not remembered well enough to see why N&E needs to be avoided in the future, and how avoiding it in the future will be good for one, then planning one's self-hypnosis it is not really an instance of rationality. If the argument is remembered, then such planning will be an instance of rationality, but it will also defeat the belief that one's plans are good ones.

Of course, chance events, or divine grace, might save the naturalist. But only chance or divine grace, not prudential rationality, could do the saving. In this way, Plantinga's argument truly is the most Calvinist of all theistic arguments. For it allows naturalists to see that, by their own lights, they are condemned to skepticism, while giving them no way to save themselves, no rational pathway over to theistic belief, not even a prudentially one.

The hopelessness of naturalists swayed by the EAAN has not been properly appreciated. In Plantinga's book (2011a), he presents the EAAN as evidence that science is in "deep discord" with naturalism, whereas much of the book argues that Christian

theism is in "deep concord" with science, contrary to popular atheistic arguments. The implication, though it is never stated, is that one should avoid beliefs that are in deep discord with science, and *should adopt* beliefs that are in deep concord with science. But for the reasons just discussed, that implication should not be drawn. Plantinga is probably aware, which is why he never says explicitly that one who wants concord with the sciences *ought to believe naturalism to be false*. But he also never acknowledges that a shift from naturalistic to anti-naturalistic belief via the EAAN would be *ir*rational.

Callum Miller, in his criticism of Law's proposed conceptual constraints on the assignment of belief contents (CC+) says the following: "Plantinga's whole point is that in order to hold on to CC+ (or something like it), one should give up E&N" (2015, 150). I could not disagree with Miller more strongly. Plantinga's point cannot be that the naturalist ought to "give up" N&E "in order to hold on to" anything, whether it be our beliefs about belief, our beliefs about R, or any beliefs whatsoever. For remember, there is no logical or probabilistic relation claimed to hold between N&E and any other belief. The relation is not alethic! So it's not as if one sees that N&E entails that something you take to be true is actually false, and therefore, N&E is likely false. Rather, believing N&E makes one unable rationally to believe things that one now takes to be true.

To give up N&E with an eye to *saving* some other belief, B, that one has, one needs rationally to believe, first, that if one did *not* believe N&E, one *could* believe B rationally, and second, one needs to downwardly adjust one's probability estimate of N&E on the basis of no new information whatsoever that bears on the probability of N&E. The first step cannot be taken, because, since the naturalist has already been

defeated in all of her beliefs, she has already been defeated in her beliefs about what would be rational if she no longer believed N&E too. Because the first step cannot be taken, there is no rational path out of the belief that N&E *in order to achieve anything else*. But even if we could set that problem aside, there still would not be an *epistemically* rational path. The second step could not be taken, (epistemically) rationally, since it would be just like our XX-pill-lottery player thinking he's won the XX-pill lottery simply because life would be better for him were he to win, which is wishful thinking. The naturalist would be shifting her credence on the basis of no relevant evidence, but merely the hope that she can believe something rationally. That would not be *epistemic* rationality.

Here, we have circled back to the opening of this paper. The EAAN is not an argument *against* naturalism. It is certainly not an argument *for* anti-naturalism. It is an argument against *rational belief in* naturalism. And none of the usual inferences one might draw from that conclusion about what a naturalist *should* do should be drawn. Thus, even if all of the criticisms of the argument I have made in this paper are misguided, and even if reflective naturalists are irrational after all, the EAAN cannot serve as one weapon in an apologetic arsenal, for instance, as part of a cumulative case for belief in God, nor function in any way as a tool for improving one's beliefs. It is just a crystal ball, really, by means of which one can foresee, to either one's horror or relief, one's epistemic fate.

Underlying philosophical issues

I promised at the outset that I would show how the EAAN hinges on (at least) four philosophical issues:

> (i) whether we have already discovered, or will discover in the near future, a compelling, independently motivated naturalized theory of content that roughly matches our pre-theoretic beliefs about belief;
>
> (ii) how our best philosophical theories, throughout their stages of development, should constrain our rational beliefs about their subject matters;
>
> (iii) how our beliefs should be shaped by higher-order evidence, and in particular, evidence about the origins of our beliefs; and
>
> (iv) whether "internal" states are in any way privileged in fending off skeptical attacks.

I think we have already seen clearly how the debate hinges on (iv), for I have used a kind of internalist assumption in deflecting looming defeaters by appealing to the differences in experiences of my own and experiences of members of the unreliable sub-populations of populations to which I believe myself to belong. If one staunchly refuses the internal any such privilege in deflection, one will not be moved by my response to the analogies.[38]

---

[38] So much the worse for one's anti-internalism, I say. Externalist: you will have a hard time convincing any audience anywhere -- with the following exceptions, perhaps: an externalism rally, and mid-mountain-climb -- that in a case where your audience believes they recently took a pill that causes, in 99% of ingesters, *hallucinations as of climbing a mountain*, and unreliability about a great many things, and in 1% of ingesters causes nothing, that they cannot rationally believe they are in the 1% that is not affected by the pill. Note also: truly hardcore externalists will oppose "defeatism" anyway. If you are reliable, they'll say, then you know.

Defeatism and Higher-Order Evidence

Let me briefly discuss the relevance of (iii). In explaining the mechanism by which naturalism defeats itself, I characterized Plantinga as first using *P(R|N&E) is low* as a partial rebutting defeater for R, and then, since it renders belief in R irrational, as an *undercutting* defeater for everything. This category of undercutting defeat is not without controversy, and there is a growing band of epistemologists who find the idea of undercutting defeat to be problematic.

Recall that the EAAN is compatible with there being excellent first-order evidence for E&N, evidence that simply raises the probability of the truth of E&N. The way Plantinga's defeat mechanism works does not bear directly on that evidence or on E&N. So, supposing there is good evidence for E&N, we have a kind of conflict, if Plantinga is right, between our first-order evidence and our higher-order evidence. The former is telling us to believe E&N. The latter is telling us to refrain from believing in E&N.

What should we do in cases of such conflict? Fans of undercutting defeaters are siding with high-order evidence. But not everyone is a fan. Anti-defeatists think that to allow "defeaters" to "undercut" one's first-order evidence is also, in a way, wrong. For it dictates that the subject does not form beliefs in the way that the first-order evidence requires. In many ways, far too complex and varied to engage here, Maria Lasonen-Aarnio, Max Baker-Hytch, Matthew Benton, Amia Srinivasan, and John Hawthorne have called defeatism into question. If they are right about undercutting defeat, the EAAN fails, for even if R is partially rebutted by P(R|N&E), an inability

rationally to believe R will not "undercut" all of one's beliefs (Lasonen-Aarnio; Baker-Hytch and Benton; Srinivasan and Hawthorne).

Undercutting is so intuitive. Why would anyone have a problem with it, in general? Here is a sample argument, much condensed and vastly oversimplified, from Lasonen-Aarnio (Lasonen-Aarnio). Consider the question of how, rationally, to adjudicate conflicts between first-order evidence and higher-order evidence. If there is an answer, it will take the form of a rule about how, rationally, to resolve such conflicts. Call this the "meta-rule". Now, for any such a meta-rule, there could conceivably be higher-order evidence against believing in it. But then, it would seem wrong simply to apply the meta-rule to adjudicate the higher-order evidence against believing the meta-rule. For, after all, the meta-rule is in question! So we need a new rule, a meta-meta-rule, to handle cases of higher-order evidence against the meta-rule. But then, this meta-meta-rule itself could be the target of higher-order evidence, and we would need yet another rule. Lasonen-Aarnio finds problems with this infinite hierarchy of rules. Suppose you bundled all such rules into one "uber-rule". Couldn't this, too, always come under higher-order fire? And wouldn't appealing to the uber-rule to adjudicate higher-order evidence against itself be inappropriate, since the rule itself is under fire? So, there is at least a prima facie difficulty here, one with which Plantinga has not dealt, and one Lasonen-Aarnio thinks stands in the way of any "defeatist" epistemology.

Here is another anti-defeatist strategy that would block Plantinga's argument. In a paper called "Religious Knowledge", Hawthorne puts forward the following two principles of knowledge:

> *Transfer* If x knows that y asserts P and x comes to believe P by trusting y with respect to P, and y knows P, then x comes to know P.
>
> *Maintenance* If x knows P by trusting y and continues to believe P on that basis, then, whatever else happens, x continues to know P. (2007, 9)

Hawthorne presents many cases in which these two principles seem to render the correct verdicts about whether one knows in the face of apparently undermining evidence For example, he writes: "Suppose a parent knows P, tells a child P and then all sorts of people tell the child that the parents have messed up, but the child sticks to his guns in believing P. Is it really so clear that the child stops knowing P?" (2007, 9)

Of course Hawthorne's principles are controversial, as are his judgments on the individual cases. But they are also not outrageous, and they are vastly *simpler* than the principles of defeat, which have yet to be articulated clearly by Plantinga or anyone else. If Hawthorne is right, and supposing one knows some proposition, p, by trusting someone, and continues to believe p despite reading Plantinga's EAAN, then by *transfer* and *maintenance*, one continues to know p, despite Plantinga's "defeater". So Plantinga's argument turns on whether Hawthorne, and others like him, are right about defeat.

The truly responsible way to address Plantinga's argument is thus with a *general theory of defeat*, relative to which the EAAN will fall out as just one case. I have not

done that here because I do not yet *have* a theory of defeat (or a "no defeat" theory either, for that matter) and I believe the EAAN can still be usefully clarified and criticized without such a theory. Nevertheless, it must be acknowledged that we are out on the periphery of this kind of epistemology here, while Hawthorne, Benton, Lasonen-Aarnio, and the others cited above, as well as their defeatist critics, are at the center.

The promise of naturalized semantics and its relevance to the EAAN

In "Content and Natural Selection", Plantinga considers three different "defeater-deflector" candidates coming from the philosophical literature on naturalized semantics: indicator semantics, functionalism, and teleosemantics (2011b). He rejects each as inadmissible, and for different reasons, and I will not get into the details of his argumentation. Nor will I present the details of the theories in question.

Naturalized semantics have, as their bases, facts about causal, nomic, or counterfactual relationships, or facts about biological relationships between creatures and features of their environments. From these bases, naturalizers of content propose theories about when a given neural state in a given environment and with a given history represents a certain proposition as true. So, for instance, a teleosemanticist notices: beavers slap their tails to signal the presence of danger; tail slaps normally cause a certain response in consuming beavers (diving under water); tail slaps come at a cost to both the producers and consumers, but is nevertheless worth that cost, in evolutionary terms. One might think these, and related facts about what beavers are signaling could be bootstrapped into an entire theory of belief content.

But building such a theory of content, not just for beaver beliefs, but for our own, is no mean feat. It is hard to see how the raw materials for naturalized semantics can distinguish the contents of different beliefs that are eternally true, or necessarily true, or merely hyperintensionally distinct. I agree with Plantinga that, for instance, Dretske's indicator semantics and Millikan's teleosemantics cannot actually account for the belief in naturalism itself.

Naturalism, if true at all, is true at all times and places. And therefore, it's hard to see how, in Dretske's sense, a neural state could "carry the information" that naturalism is true (Plantinga, 2011b, 450), and it is hard to see how naturalism and the belief in naturalism could have played the appropriate sorts of roles in an evolving signaling system with a producer, a shared representation token used by consumers, triggering "normal" and evolutionarily advantageous responses on the part of the consumers. Unlike a beaver's tail slap, there is no obviously "normal" or adaptive response to the belief that naturalism is true. As Plantinga dryly observes, "It is only the occasional member of the Young Atheist's Club whose reproductive prospects will be enhanced by proclaiming naturalism" (2011b, 457).

Plantinga does not claim to have addressed *all* naturalized semantics, of course, but hopes he has done some damage to the very idea, and that we can see how a candidate deflector on the basis of a naturalized semantics will probably not be admissible to deflect the defeater of *P(R|N&E) is low* (2011b, 445, 458).

Now, it is not clear dialectically who has the argumentative burden here. I think from Plantinga's perspective, if one thinks there is a connection between belief and truth,

one is welcome to trot out one's theory of belief, one's naturalized semantics, as a potential defeater-deflector. He does not think theories of the nature of belief itself are question begging: "Considered beliefs about the nature of belief itself can, presumably, be properly added…" (2011b, 440). Yet, one is not allowed to add something as simple as "true belief is adaptive" (2011a, 335). That is information about the actual world, not the nature of belief. I puzzled over the difference in these two judgments at length earlier in the paper, noting that beliefs of the second sort are used to evaluate theories of the first sort, and I never found a principled reason why Plantinga makes those two different judgments.

In any case, what I want to point out here is that there are two responses to Plantinga's criticisms of the three naturalized theories of content he surveys in "Content and Natural Selection". The first response is to develop a naturalized semantics, in full detail, that avoids his criticisms. This, for instance, is the response of Brian Leahy, who defends teleosemantics at length from Plantinga's objections (2013). Leahy's response is only as good, though, as the specific version of teleosemantics he defends. And teleosemantics, like any reductive philosophical theory of anything outside of mathematics, has its share of outstanding issues (see Neander for the relevant challenges to teleosemantics). In fact, every naturalized theory of content has its share of issues.

Another kind of response points out that we are at the very beginning of the project of naturalizing content. We should not demand completed theories at this stage, nor should we tie our beliefs about belief to the consequences of our best theories of belief at present. Nevertheless, we might hold onto the idea that a true naturalized theory

of content is either (a) nearing discovery, or (b) true, whether or not it will soon be discovered.

For comparison, think about the naturalist on consciousness. Most naturalists acknowledge that we do not yet have an adequate material explanation of conciousness. But what naturalists do not (and should not) do is to give up the idea that they themselves are conscious. Yet, why don't they? On their best material theories, there is no explanatory link from material states to conscious states! And they believe everything is material! Another alternative would be to give up on materialism itself. Yet many materialists also refuse to do this, even while acknowledging the explanatory gap.

How is the materialist permitted to do this? Well, we acknowledge that theorizing is hard. We may have good reason to be materialists and also have good reason to think we are conscious, and yet, according to our best theories of the material, theories limited to material and topic-neutral vocabulary only, none of us is conscious, since our material theories cannot explain the phenomenon of consciousness whatsoever. Some materialists think that if we work hard enough, we will find that elusive theory of consciousness. Others, like Colin McGinn, think that by our nature, we are barred from ever understanding such a theory, though one must be true (McGinn).

A parallel situation holds for the naturalizer of content. Naturalistic theories generally make true belief adaptive. They have problems accounting for mathematical beliefs, philosophical beliefs, and so on. But a naturalizer, one who has a certain "favorite" or "best" naturalized theory of content, is not forced to adopt the view that *she herself has no abstract mathematical or philosophical beliefs*. Rather, she can think that

her theory just hasn't captured those beliefs *yet*, but it will in time, as it is further developed. Or, she can adopt the McGinn-like posture and suppose that no one will ever understand how such beliefs are possible, given that we are material beings, yet still we do have such beliefs, and still such beliefs' contents are determined in material ways.

Now, here is where the dialectic is genuinely confusing. Suppose one thinks that a naturalized theory of content, on which true belief is generally adaptive, is either on the horizon, or is not on the horizon, but is nevertheless true. Is such a belief an *admissible* defeater deflector for the looming defeater of *P(R|N&E) is low*? I'm sure Plantinga would argue that it is not, and that one has simply begged the question. Yet, why isn't the right attitude to take towards reductive theories that they are probably incomplete? Why is one bound to believe the contents of one's best theory, whilst knowing that reductive theories are, almost every single one of them, flawed? And here again, the objector to Plantinga will urge that she is attacking premise 1, not offering a deflector to a "looming" defeater. What she thinks is that, once belief is properly understood, P(R|N&E) will not be seen to be low.

I will not go back to the debate about whether this is indeed a problem for premise 2 or for premise 1, and I will leave these question unsettled. But note that if one, as a general rule, does not think one's beliefs on a topic are bound by one's current best theories on that topic, then one need not actually provide a counterexample-free naturalized semantic theory on which true belief is adaptive in order to believe, rationally, that true belief is adaptive. And whether we locate the belief that true belief is

adaptive as a defeater deflector or as an objection to premise 1, it is a problem for the EAAN.

Summary of how the EAAN rests on broader philosophical issues

I take myself to have shown how the EAAN depends on these further philosophical issues: the problem of high-order evidence; the privilege, if any, of our beliefs about our internal states, in fending off skeptical attacks; the prospects of a naturalized semantics that preserves our commonsense beliefs about belief; and the appropriate attitudes towards fledgling reductive theories, for a reductivist. I have made that case in terms of what assumptions about each of these issues would sink the EAAN. Let me, in this summary, say instead what assumptions the EAAN requires.

On the issue of internalism, Plantinga needs to deny that beliefs about our internal states can play any special role in fending off the threats of looming defeaters.

On the issue of how conflicts between higher-order evidence and first-order evidence should be resolved, Plantinga needs it to be the case that higher-order evidence of one's unreliability trumps first-order reasons for belief. Plantinga, obviously, stands on the side of the "defeatists".

On the issue of naturalized theories of content, Plantinga needs it to be the case that there are no independently plausible naturalistic theories of content according to which: (i) adaptive beliefs are more likely true than false, and (ii) naturalism and evolution themselves are possible to believe. He needs it to be the case not only that

none of the known theories fits that bill, but that it is implausible that there could be a rational hope for such a theory. Moreover, he needs it to be the case that it would be irrational now to believe that such a theory could be true, even if we have not yet discovered it, or may never discover and develop it, due to our cognitive limitations. Instead, Plantinga needs it to be the case that we are bound to believe the consequences of our best theories at the time, however counterintuitive those consequences may be.

If one sides with Plantinga on each of these issues, then, for one, the argument will still stand a chance of success, provided one can answer the many objections and requests for clarification posed throughout this paper.

Conclusion

In the early sections of this paper, I tried to get clear on exactly what the EAAN is supposed to be, in its latest and most potent forms. This was no easy task, and throughout the paper I traced a number possible developments of the argument and exposed ambiguities in the premises. I explored the difficulties in trying to object to premise 1 at all. Plantinga, it seems, has tailored the rules of engagement such that any objection to premise 1 is properly treated only as an objection to premise 2, which then has built into itself a way of disqualifying criticisms as question-begging.

With what I took to be the strongest and most contemporary form of the argument in hand, I critically examined premise 2, the defeat premise. I showed that Plantinga's notion of a "defeater-deflector" on the one hand concedes too much to the skeptic, since every belief is a "looming defeater" for every other belief, on my reading. Nevertheless,

on the other hand, I found that Plantinga's *explicit* criteria for admissible deflection are still easy enough to meet for a naturalist, who can "deflect" the "looming defeater" Plantinga poses to the naturalist. (I offered several examples of deflectors that meet Plantinga's explicit criteria.) Given the incompleteness of Plantinga's criteria for deflection, his entire argument rests on certain analogous cases. I showed that those analogies are *ambiguous*, and however the ambiguity is resolved, the naturalist has a hopeful response.

Lastly, I made two meta-remarks about the EAAN. First, I pointed out that it cannot, even if it is sound, "save" a naturalist from skepticism, or play any role in such a saving effort: the EAAN rules out any prudentially or epistemically rational pathway from naturalistic belief to rational belief in anything. Second, I outlined some of the broader philosophical issues underlying the argument -- the prospects for a sane naturalistic theory of content, the constraints imposed on rational belief by the contents of our best fledgling theories, general issues of higher-order evidence, and whether the internal can play a privileged role in fending off skeptical attacks. I argued that the EAAN presupposes substantive, controversial positions on each.

The EAAN appears to be a straightforward argument against naturalistic belief, an argument that must, one thinks, rely on some simple sophistical trick. That appearance is misleading. I could not "refute" the argument. It was hard enough to figure out exactly what the argument was. But neither could I find, among the many possible versions of the EAAN I constructed, anything that would convince a fair-minded

naturalist. That does not mean a compelling version of the EAAN *cannot* be given, but it does mean that no such version *has yet* been given.

Works Cited

Baker, Lynne. *Saving Belief*. Princeton: Princeton University Press, 1987.

Baker-Hytch, Max, and Matthew Benton. "Defeatism Defeated." *Philosophical Perspectives* 29.1 (2015): 40–66.

Block, Ned. "Inverted Earth." *Philosophical Perspectives* 4 (1990): 53–79.

Boghossian, Paul. "The Status of Content." *The Philosophical Review* 99.2 (1990): 157–184.

Bourget, David, and David Chalmers. "What Do Philosophers Believe?" *Philosophical Studies* 170 (2014): 465–500.

Childers, Geoff. "What's Wrong with the Evolutionary Argument Against Naturalism?" *International Journal for Philosophy of Religion* 69.3 (2010): 193–204.

Churchland, Patricia. "Epistemology in the Age of Neuroscience." *Journal of Philosophy* 84 (1987): 546–553.

Cross, Troy. "Skeptical Success." *Oxford Studies in Epistemology*. Ed. John Hawthorne and Tamar Szabo Gendler. Vol. 3. Oxford: Oxford University Press, 2010. 35–62.

Fitch, Frederic. "A Logical Analysis of Some Value Concepts." *The Journal of Symbolic Logic* 28.2 (1963): 135–142.

Fitelson, Branden, and Elliott Sober. “Plantinga’s Probability Arguments Against Evolutionary Naturalism.” *Pacific Philosophical Quarterly* 79.2 (1998): 115–129.

Fodor, Jerry. “Is Science Biologically Possible?” *Naturalism Defeated?*. Ed. James Beilby. Ithaca and London: Cornell University Press, 2002. 30–42.

Goodman, Nelson. *Fact, Fiction, and Forecast*. 4th ed. Cambridge, MA: Harvard University Press, 1983.

Hawthorne, John. “Advice for Physicalists.” *Philosophical Studies* 109.1 17–52.

---. “Religious Knowledge.” *Philosophic Exchange* 37.1 (2007): 1–11.

Hawthorne, John, and Amia Srinivasan. “Disagreement Without Transparency: Some Bleak Thoughts.” *The Epistemology of Disagreement: New Essays*. Ed. David Christensen and Jennifer Lackey. Oxford: Oxford University Press, 2013. 9–30.

Kvanvig, Jonathan. “Two Approaches to Epistemic Defeat.” *Alvin Plantinga*. Ed. Deane-Peter Baker. Cambridge, England: Cambridge University Press, 2007. 107–124.

Lasonen-Aarnio, Maria. “Higher-Order Evidence and the Limits of Defeat.” *Philosophy and Phenomenological Research* 88.2 (2014): 314–345.

Law, Stephen. “Naturalism, Evolution, and True Belief.” *Analysis* 72.1 (2012): 41–48.

---. “Plantinga’s Belief-Cum-Desire Argument Refuted.” *Religious Studies* 47.2 (2011): 245–256.

Leahy, Brian. “Can Teleosemantics Deflect the EAAN?” *Philosophia* 41 (2013): 221–238.

Mayo Clinic Staff. “Creutzfeldt-Jakob Disease.” *Mayo Clinic*. 2015. Web. 6 July 2016.

McGinn, Colin. “The Problem of Philosophy.” 1993.

McLaughlin, Brian. “Type Epiphenomenalism.” *Philosophical Perspectives* 3 (1989): 109–135.

Miller, Calum. “Response to Stephen Law on the Evolutionary Argument Against Naturalism.” *Philosophia* 43.1 (2015): 147–152.

Mirza, Omar. “A User’s Guide to the Evolutionary Argument Against Naturalism.” *Philosophical Studies* 141.2 (2008): 125–146.

---. “The Evolutionary Argument Against Naturalism.” *Philosophy Compass* 6.1 (2011): 78–89.

Nagel, Thomas. *Mind and Cosmos*. Oxford: Oxford University Press, 2012.

Nathan, N.M.L. “Naturalism and Self-Defeat: Plantinga’s Version.” *Religious Studies* 33.2 (1997): 135–142.

Neander, Karen. “Teleological Theories of Mental Content.” *The Stanford Encyclopedia of Philosophy*. Ed. 2012.

O’Connor, Timothy. “A House Divided Against Itself Cannot Stand.” *Naturalism Defeated?*. Ed. James Beilby. Ithaca and London: Cornell University Press, 2002. 129–134.

Otte, Richard. “Conditional Probabilities in Plantinga’s Argument.” *Naturalism Defeated?*. Ed. James Beilby. Oxford: Oxford University Press, 2002. 135–152.

Pew Research Center. *Public Praises Science; Scientists Fault Public, Media*. Washington, D.C., 2009.

Pew Research Center. *How Different Groups Think About Scientific Issues*. Washington, D.C., 2014.

Plantinga, Alvin. “An Evolutionary Argument Against Naturalism.” Veritas Forum. University of Southern California. Lecture.

---. “An Evolutionary Argument Against Naturalism.” *Logos.* 12 (1991): 27–48.

---. “Content and Natural Selection.” *Philosophy and Phenomenological Research* 83.2 (2011): 435–458.

---. “Introduction: The Evolutionary Argument Against Naturalism.” *Naturalism Defeated?*. Ed. James Beilby. Ithaca and London: Cornell University Press, 2002. 1–14. Print.

---. “Reply to Beilby’s Cohorts.” *Naturalism Defeated?*. Ed. James Beilby. Ithaca and London: Cornell University Press, 2002. 204–275.

---. *Warrant and Proper Function*. Oxford: Oxford University Press, 1993.

---. *Warranted Christian Belief*. Oxford: Oxford University Press, 2000.

---. *Where the Conflict Really Lies: Science, Religion, and Naturalism*. Oxford: Oxford University Press, 2011.

---. 'Evolution vs. Atheism'. Veritas Forum. 2013. Lecture.

---. 'Science and Religion: Where the Real Conflict Lies'. Veritas Forum. 2009. Lecture.

---. 'Science and Religion: Why Does the Debate Continue?' Veritas Forum. 2013. Lecture.

Plantinga, Alvin, and Stephen Law, *Alvin Plantinga versus Stephen Law on the Evolutionary Argument Against Naturalism*. Unbelievable? 2010. Podcast.

Prasetya, Yunus. "An Analysis of Stephen Law's Objection to Alvin Plantinga's Evolutionary Argument." *Polymath: An Interdisciplinary Arts & Sciences Journal* 4.3 (2014): 22–26.

Quine, Willard Van Orman. *From a Logical Point of View: Nine Logico-Philosophical Essays, Second Revised Edition*. Revised edition. Cambridge, Mass: Harvard University Press, 1980.

Reppert, Victor. "Eliminative Materialism, Cognitive Suicide, and Begging the Question." *Metaphilosophy* 23.4 (1992): 378–392.

Robb, David, and John Heil. "Mental Causation." *Stanford Encyclopedia of Philosophy*. 2013.

Schillp, Paul. *The Philosophy of G.E. Moore*. New York: Tudor Publishing Company, 1942.

Stich, Stephen. *The Fragmentation of Reason*. Cambridge, MA: MIT Press, 1993.

Talbott, W.J. "The Illusion of Defeat." *Naturalism Defeated?*. Ed. James Beilby. Ithaca and London: Cornell University Press, 2002. 153–164.

Trivers, Robert. *Deceit and Self-Deception: Fooling Yourself the Better to Fool Others*. London: Allen Lane, 2011.

Ye, Feng. “Naturalized Truth and Plantinga’s Evolutionary Argument Against Naturalism.” *International Journal for Philosophy of Religion* 70.1 (2011): 27–46.