Why AI Doomsayers are like Sceptical Theists and Why it Matters

John Danaher*

Forthcoming in ***Minds and Machines*** DOI: 10.1007/s11023-015-9365-y

Abstract

An advanced artificial intelligence (a "superintelligence") could pose a significant existential risk to humanity. Several research institutes have been set-up to address those risks. And there is an increasing number of academic publications analysing and evaluating their seriousness. Nick Bostrom's Superintelligence: Paths, Dangers, Strategies represents the apotheosis of this trend. In this article, I argue that in defending the credibility of AI risk, Bostrom makes an epistemic move that is analogous to one made by so-called sceptical theists in the debate about the existence of God. And while this analogy is interesting in its own right, what is more interesting is its potential implication. It has been repeatedly argued that sceptical theism has devastating effects on our beliefs and practices. Could it be that AI-doomsaying has similar effects? I argue that it could. Specifically, and somewhat paradoxically, I argue that it could lead to either a reductio of the doomsayers position, or an important and additional reason to join their cause. I use this paradox to suggest that the modal standards for argument in the superintelligence debate need to be addressed.

Keywords: Superintelligence; Artificial General Intelligence; AI risk; Existential Risk; Sceptical theism

* PhD; Lecturer in Law, NUI Galway, Ireland

1. Introduction

An advanced artificial intelligence (a "superintelligence") may pose an existential threat to humanity. Such an intelligence would be significantly better than us at accomplishing its goals, and its goals may be antithetical to our own.[1] To use the now-classic example, an advanced artificial intelligence might be a *paperclip maximiser*: an entity dedicated to using its intelligence to create the maximum possible number of paperclips, no matter how many humans have to suffer in the process. As one famous AI-risk advocate puts it, "[such an] AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else" (Yudkowsky 2008).

Proponents of AI doomsday arguments of this sort have been gaining traction in recent years. Several research centres now dedicate some or all of their resources to addressing the risks that a superintelligence could pose.[2] And the number of academic works analysing those risks is also on the rise.[3] Nick Bostrom's book, *Superintelligence: Paths, Dangers, Strategies*, represents the apotheosis of this trend. It is a sophisticated, carefully argued, single-volume assessment of AI-risk, written by one of the leading figures in the field.

I think there is much of merit in Bostrom's work, and much that deserves careful critical discussion. In this article I want to hone-in on one aspect of it. In defending the credibility of AI-doomsday scenarios, and in criticising those of a more sunny disposition, Bostrom makes a certain epistemic move that is directly analogous to one made by sceptical theists in the debate about God's existence. My goal is to highlight and explore the implications of this move.

---

[1] Here I appeal to two theses defended by Bostrom in his recent book *Superintelligence* (Bostrom, 2014): the strategic advantage thesis and the orthogonality thesis. The latter thesis is particularly important for the doomsday scenario discussed in the text. It maintains that pretty much any level of intelligence is compatible with pretty much any final goal. The thesis has been defended elsewhere as well (Bostrom, 2012; Armstrong, S, 2013).

[2] The three leading examples are *The Future of Humanity Institute* based at Oxford University and headed by Nick Bostrom (see http://www.fhi.ox.ac.uk); the *Centre for the Study of Existential Risk* or CSER, based at Cambridge University (see http://cser.org); and the *Machine Intelligence Research Institute* or MIRI, not affiliated to any university but based out of Berkeley, CA (see http://intelligence.org). Only the latter dedicates itself entirely to the topic of AI risk. The other institutes address other potential risks as well.

[3] In addition to Bostrom's work, which is discussed at length below, there have been Eden, Moor, Soraker and Steinhart (2012); Blackford and Broderick (eds) (2014); Chalmers, D. (2010), which led to a subsequent symposium double-edition of the same journal, see *Journal of Consciousness Studies* Volume 19, Issues 1&2

To be more precise, my goal is to defend three major claims. First, I want to argue that the analogy is direct and real: not merely an illusion or a mis-reading of Bostrom's work. Second, I want to point out that the move made by sceptical theists is commonly thought to have significant epistemic and practical consequences (in this respect I will call upon arguments that I and others have defended in the past). And third, I want to argue that the move made by Bostrom may have similar consequences. Specifically, and somewhat paradoxically, I argue that it may lead to a *reductio* of his position, or supply an *a fortiori* endorsement of it. Either way, the analogy has significant implications for the ongoing debate about AI risk, particularly for the modal standards within that debate (i.e. the standard for making claims about what is possible or probable for a superintelligence).

Before I begin this, however, I must address a general point. I am well aware that others ridicule the beliefs of AI doomsayers (and utopians) on the grounds that there is something quaintly religious about their beliefs. There is, for example, a consistent trope in popular culture whereby the "singularity" is referred to as the "Rapture of the nerds" (Doctorow & Stross 2012). And there are academic commentators who critique singularitarians on the grounds that their belief system is *fideistic*, *i.e*. faith-based or lacking reasonable foundation (Bringsjord et al 2012). Although the argument I present draws upon similar analogies between AI doomsayers and religious believers, it differs from these critiques in at least two important ways. First, I write from a position of sympathy: I don't wish to ridicule or pour scorn on the arguments that have been developed by AI doomsayers; on the contrary, if anything my goal is to "debug" the argumentative framework in order to make it stronger. Second, my argument is quite narrow. I grant many of the claims presented by AI-doomsayers, and focus-in on a particular epistemic move. Legitimate concerns about AI risk can easily survive my criticisms.

2. Sceptical Theism and Bostrom's Treacherous Turn

To appreciate the argument I am presenting, some familiarity with the literature on the problem of evil and the sceptical theist response to it are necessary. So let's start by filling in some of the required detail. The problem of evil is the classic atheological argument. It comes in two major forms. Logical versions of the problem claim that

God's existence is impossible given the existence of evil; evidential versions claim that God's existence is improbable given the existence of evil. Central to the discussion is the claim that *gratuitous evil* (in its many instances) provides evidence against the existence of God.[4] Gratuitous evil is evil that is not logically required for some greater, outweighing good. The focus on gratuitous evil makes sense given that God is classically conceived of as an omnibenevolent, omnipotent and omniscient being. The argument is that a being with those properties would not allow evil to occur unless it was absolutely necessary for some greater good.[5]

The difficulty lies in establishing that there really are gratuitous evils. Defenders of the evidential problem will argue that we can infer the *likely* existence of such evils from their *seeming* existence. In other words, they will argue that we are entitled to infer the existence of *actually gratuitous evils* from the existence of *seemingly gratuitous evils*. The latter are evils that do not appear to be logically necessary for a greater good. There are many examples in the literature. Rowe, famously, gives the hypothetical of a deer suffering in a forest fire with no one around to see it; Maitzen uses the case of Dominick Calhoun, a four year-old who died after being beaten and burned by his mother's boyfriend (Maitzen 2013); and Street uses a case involving a car accident that led to the decapitation of a six year-old girl (Street, forthcoming). Each of these authors argue that since the suffering inflicted in these cases does not seem to be necessary for a greater good, it probably isn't necessary for a greater good.

Sceptical theists deny the legitimacy of this inference.[6] They argue that you cannot go from the seemingly gratuitous nature of evil to its actually gratuitous nature. This is because there may be *beyond-our-ken reasons* for allowing such evils to take place. We are cognitively limited human beings. We know only a little of the deep structure of reality, and the true nature of morality. It is epistemically possible that the seemingly gratuitous evils we perceive are necessary for some greater good. Michael Bergmann develops this sceptical posture into a series of four principles. These principles can be summarised as follows:[7]

[4] The standard presentation is that of William Rowe (1979); for a more detailed overview, see Trakakis, N. (2007).
[5] I defend this argument from recent attacks on the "logical necessity" condition in Danaher (2014)
[6] The idea was introduced originally by Wykstra, S. (1996). For more up-to-date overviews, see McBrayer, J (2010); Dougherty, T. (2012); Dougherty and McBrayer (eds) (2014).
[7] The summary is based on the discussion of sceptical theism in Bergmann, M. (2001) and Bergmann, M. (2009).

> We have no good reason for thinking that the possible goods and evils (and entailment relations between those goods and evils) of which we know, are representative of all the possible goods and evils (and entailment relations between them) that there are.

God, on the other hand, would have access to all this information. He would know what we do not.

Several features of sceptical theism should be noted. First, note how it is introduced to block a standard objection to the theistic worldview. Second, note how its plausibility depends upon claims about our nature, God's nature and the relationship between the two: it works because God is cognitively supreme and we are cognitively limited. Third, and finally, note how it tries to stop us from making certain standard inductive inferences, in this case inferences from the seemingly gratuitous nature of the suffering to their actually gratuitous nature. My claim is that these features have their analogues in the arguments of AI-doomsayers.

So let's see what the AI-doomsayers have to say. Academic discussion of AI risk has been much more limited than academic discussion of the problem of evil. Consequently, it would be wrong to suggest that there is a consensus view on the arguments and concepts that shape the debate. I cannot hope to review what everyone has written on this topic,[8] so instead I will focus on the arguments presented in Nick Bostrom's book *Superintelligence*. I justify this selective approach on the grounds that Bostrom's book provides the most comprehensive treatment of the topic to date.

In the book, Bostrom develops a three-premised argument for the doomsday scenario. I will describe them here without critical evaluation. My purpose is not to critique the premises of the argument, but rather its implications. That's not to say that I agree entirely with the premises – I think there are plausible criticisms to be made. It is simply to say that my present interests lie elsewhere.

The first premise claims that the first superintelligence to be created — in virtue of being first — could obtain a decisive strategic advantage over all other forms

[8] See fn 3 above for sources.

intelligence (including human forms).[9] This decisive advantage would allow the superintelligence to take control and shape the future of all intelligent life on earth. That might be fine if the superintelligence were benevolent but, according to Bostrom, there is no reason to think that it would be. This is where the second and third premises of his argument come in. The second premise points out that there is no necessary relationship between intelligence and benevolence. Pretty much any level of intelligence is compatible with pretty much any final goal.[10] This is what gives rise to examples like that of the superintelligent paperclip maximiser, mentioned in the introduction.[11] This claim is then added to the third premise of the argument, which points out that although a superintelligence could have pretty much any final goal, it is likely to converge upon certain instrumental goals that are inimical to human interests.[12] For example, it is likely to engage in extensive resource acquisition, which could include the acquisition of human bodies or other resources upon which humans rely.

This gives us Bostrom's argument for doom. As he puts it:

> *"[T]he first superintelligence may [have the power] to shape the future of Earth-originating life, could easily have non-anthropomorphic final goals, and would likely have instrumental reasons to pursue open-ended resource acquisition. If we now reflect that human beings consist of useful resources…and that we depend on many more local resources, we can see that the outcome could easily be one in which humanity quickly becomes extinct."*
>
> (Bostrom 2014, 116)

The implication of this doomsday argument is straightforward. Anyone who is planning to create or design the architecture for an artificial general intelligence (AGI) faces a significant "control problem" (Bostrom 2014, ch. 9). The original engineers and designers will, no doubt, have a certain set of goals they have in mind for the AGI.

[9] The argument for this is found in chapter 5 of Bostrom's book.
[10] This is the orthogonality thesis as defended in Bostrom (2012) and Armstrong (2013).
[11] This orthogonality thesis could be criticised. Some would argue that intelligence and benevolence go hand in hand, i.e. the more intelligent someone is the more likely they are to behave in a morally appropriate manner. I have some sympathy for this view. I believe that if there are genuine objectively verifiable moral truths, then the more intelligent the more likely they are to discover and act upon the moral truth. Indeed, this view is popular among some theists. For instance, Richard Swinburne has argued that omniscience may imply omnibenevolence. I am indebted to an anonymous reviewer for urging me to clarify this point.
[12] This is the instrumental convergence thesis. See Bostrom (2012) and Bostrom (2014), pp. 109-114

Those goals might be perfectly compatible with human flourishing.[13] The danger is that they will not be able to control the AGI once it crosses a certain threshold of intelligence. It will gain the upper hand and be able to "shape the future of Earth-originating life" in an existentially threatening way.

Is there any way to avoid these dangers? Here's where the critics of the doomsday scenario come in.[14] They suggest an obvious riposte to Bostrom's argument: just test any AGI for benevolence before releasing it into the wild. The idea would be to keep the AGI in a safe, controlled environment (referred to in the literature as a "box")[15] and repeatedly test it for its benevolence and safety. The "box" in question would need to contain a simulated environment that allows us to recreate real-world situations and scenarios. When we have done enough trials testing the AGI in that experimental environment, and we are happy that the AGI *seems* to pose no risk to human life, we can release it from the "box".

We can call this the "empirical testing" objection to the doomsday scenario. Note how it is based on an induction from empirical evidence. It says that if it *seems* like the AGI is behaving benevolently after multiple, repeated empirical testing in an ecologically valid experimental environment, then we are warranted in believing *that it is* benevolent. The pattern of reasoning here is practically identical to that used by proponents of the evidential argument from evil. Once again, we go from what seems to be the case to what is actually the case.

And once again the defenders of the original argument try to block that inference. Bostrom does this by introducing a concept he calls the "Treacherous Turn" (Bostrom 2014, 116-119). AGIs, like humans, would be strategic actors: they would adopt behavioural strategies that plan for and anticipate likely human responses. We know from everyday human-to-human interactions that a common strategy for achieving one's goals includes being deceptive towards others. For example, I may feign a lack of interest in my colleague's cutting-edge research, in the hope that he will trust me and disclose all its details to me, then, knowing that he struggles to finish his writing

[13] They may not be if the designers themselves of malevolent goals, but that's a distinct issue, having to do with our understanding of human agency, not superintelligent machine agency.
[14] The leading critics in the academic literature are probably Ben Goertzel and Richard Loosemore; online, Alexander Kruel maintains a regularly-updated blog critiquing the doomsday scenario. See http//kruel.co
[15] This is how Bostrom (2014) describes is at pp. 129-131; it is also referred to as 'Leakproofing' by Yampolskiy (2012).

projects, I could steal his ideas and present them as my own, thereby claiming all the glory for myself. Couldn't AGIs do the same thing? Couldn't they realise that we are testing them with a view to determining whether they are safe or "friendly"? And couldn't they then adopt a strategy of playing nice in order to get out of the box?

Bostrom certainly seems to think they could:

> *"An unfriendly AI of sufficient intelligence [may realise] that its unfriendly goals will be best realized if it behaves in a friendly manner initially, so that it will be let out of the box. It will only start behaving in a way that reveals its unfriendly nature when it no longer matters whether we find out; that is, when the AI is strong enough that human opposition is ineffectual."*
>
> (Bostrom 2014, 117)

This is the essence of the treacherous turn: An AGI may *seem* to be perfectly benevolent, cooperating with us and protecting our interests, but it may *actually* have long-term plans that are inimical to our interests. It may, for all we know, take a treacherous turn after lulling us into a false sense of security. The inference that the critics are trying to make is blocked by the possibility of the AGI taking a treacherous turn. The principle applies specifically to the case of inductive inferences about the AGI and not to all similar types of inductive inference, e.g. inductive inferences about the potential deceptiveness of one's human friends. This is because the AGI is, in the view of Bostrom and others, capable of playing a much 'longer' and much more deceptive game than the typical human.

You may wonder how seriously we should take this possibility and whether Bostrom really intends for it to have the same effect as sceptical theism is intended to have. Sceptical theists want their claims to completely rule out any attempt to resurrect the evidential problem of evil. They claim that we just do not have the cognitive resources that allow us to make inferences from the seemingly gratuitous nature of an evil to its actually gratuitous nature. But surely Bostrom isn't claiming the same is true for seemingly benign acts performed by an AGI? Surely there is some level of empirical evidence that would satisfy him that an AGI posed no risk to humans?

Maybe, but it's not obvious from his work, and there are at least four reasons for thinking he intends for it to have a deep effect, an effect that will permanently alter our attitude toward the creation of machine superintelligence. First, there is the simple fact that he brings it up several times as an objection to various proposals for solving the control problem (Bostrom 2014, chs 8 & 9). Second, there is the fact that he doubts whether the problem would be obvious enough to be factored into the design process.[16] He thinks that we may be overconfident in making inferences from past experience with the design of artificial intelligence. Third, there is the fact that he explicitly warns us against interpreting the idea "too narrowly" (Bostrom 2014, 119). In doing so, he points out that an AGI could deceive us in a variety of ways, many of which could involve long-term thinking that is beyond our ability to fully comprehend. So, for example, the AGI could deliberately pretend to have fewer capabilities than it actually has;[17] or it could deliberately encourage its own destruction because it has predicted that this will lead the engineers to create another, similar AGI with a similar utility function;[18] or it could even take advantage of the fact that we may not be able to realize what is or is not in our long-term interest.[19] Fourth, and finally, there is the fact that this deeper interpretation is consistent with Bostrom's approach to existential risk. As Bostrom defines it (Bostrom 2013), an existential risk something that threatens the extinction or fundamental well-being of intelligent life on earth. It is something that has potentially "crushing" or "hellish" impacts and a "pan-generational" reach (Bostrom 2013, 17). His doomsaying argument suggests that a superintelligent AI may pose such a risk to humanity. But since the effects are so potentially devastating we should treat them seriously, even if their overall probability is low.[20] Thus, the treacherous turn possibility should not be taken lightly, even if the evidence weighs against it.

By now, the analogy between the treacherous turn and sceptical theism should be apparent. In both instances we have an initial belief of great significance (the existence

[16] Bostrom (2014), p. 117 "One might think that the reasoning described above is so obvious that no credible project to develop artificial general intelligence could possibly overlook it. But one should not be too confident that this is so." He then proceeds to give an example which suggests we may be overconfident in our inferences from past experiences.
[17] Bostrom (2014), p. 117 "an unfriendly AI may become smart enough to realize that it is better off concealing some of its capability gains." This could even involve adjusting its source code to deceive the testers.
[18] Bostrom (2014), p. 119 "For example, an AI might not play nice in order that *it* be allowed to survive and prosper. Instead, the AI might calculate that if it is terminated, the programmers who built it will develop a new and somewhat different AI architecture, but one that will be given a similar utility function."
[19] Bostrom (2014), p. 113 and later at chapter 12 and the discussion of the value-loading problem.
[20] To be clear, this does not mean that an infinitesimal probability of an existential risk should be taken seriously. But, say, a .05 or .1 risk may be sufficient, given what is at stake.

of god in the first case and the possibility of an AI doomsday scenario in the second). These beliefs concern the existence of superpowerful and superintelligent agents (God or the Superintelligent AGI). The beliefs are challenged by certain objections (the evidential problem of evil and the empirical testing objection). Both of these objections are premised on the notion that we can draw reliable (albeit probabilistic) inferences from the seemingly benign/evil nature of event to their actual natures. These objections are rejected by holders of the original belief. In both instances the believers block the inference from "seeming" to "actual" by appealing to our cognitive limitations and the qualities of the superintelligence: the superintelligence is in it for the long haul, we may not fully appreciate how what we observe of its behaviour connects up to its other (ineffable?) final goals. There are only three real differences between the cases. The first is that God isn't simply *superintelligent* and *superpowerful*; he is maximally intelligent and maximally powerful. The second is that God is usually supposed to have created the entire universe and so every aspect of that creation speaks (in an inductive sense) to his nature; the same is not true of a superintelligent AI, though it may (possibly?) eventually acquire complete control over the universe. The third is that in the case of sceptical theism the inference being blocked is from the seemingly evil nature of an event, whereas in the case of Bostrom it is from the seemingly benign nature of an event. These differences, however, don't erase the underlying similarities.

I believe these similarities are interesting in their own right, but not overly significant. After all, similar argumentative strategies are adopted in many areas of philosophy. Their significance lies in their practical and epistemic consequences. If the consequences for the treacherous turn end up being analogous to the consequences for sceptical theism, then it may amount to something more than a mere intellectual curio.

3. The Practical and Epistemic Costs of Sceptical Theism

To develop the argument further we need to first consider the practical and epistemic implications of sceptical theism. Sceptical theists have to walk a very fine line. They have to use their sceptical posture to deny the inductive inference used by proponents of the evidential problem of evil, while at the same time ensuring that the sceptical posture goes no further than that, *i.e.* that it doesn't affect other beliefs or

arguments to which they are committed. As Lovering puts it, they need to ensure that theirs is a "narrow" form of sceptical theism (Lovering 2009).

The problem is that it is not clear that they can do this. As many critics have pointed out, a commitment to sceptical theism would seem to have several unwelcome consequences.[21] Lovering perhaps goes furthest by arguing that a commitment to sceptical theism undermines every other theistic argument. The reason for this is that sceptical theism calls into question our ability to know what God would or would not do, and every argument for God's existence relies, implicitly or explicitly, on a claim about what God would or would not do (Lovering 2009). Other authors have made similar critiques, albeit with a narrower focus. For example, Wielenberg (2010; 2014) argues that sceptical theism casts into doubt our ability to tell whether or not God is lying to us; Hasker (2010) and Piper (2008) argue that it undercuts a number of standard inductive inferences; and Sehon (2010), Maitzen (2013), Oppy and Almeida (2003), and [reference omitted] have all argued that it leads to moral uncertainty/paralysis. These latter two objections are particularly interesting because they suggest that sceptical theism has implications outside of the religious domain. The commitment to sceptical theism has a "contaminating" effect on an entire ecosystem of beliefs and practices [reference omitted].

To illustrate the problem, it is worth exploring the moral uncertainty/paralysis critique in more detail. The claim made by proponents of that critique is that anyone who believes in God and who uses sceptical theism to resist the problem of evil, confronts a dilemma whenever they are confronted with an instance of great suffering. Suppose you are walking through the woods one day and you come across a small child, tied to tree, bleeding profusely, clearly in agony, and soon to die. Should you intervene, release the child and call an ambulance? Or should you do nothing? The answer would seem obvious to most: you should intervene. But is it obvious to the sceptical theist? They assume that we don't know what the ultimate moral implications of such suffering really is. For all they know, the suffering of the child could be logically necessary for some greater good, (or not, as the case may be). This leads to a practical dilemma: either they intervene, and possibly (for all they know), frustrate some greater good; or they do nothing and possibly (for all they know) permit some great evil. As Sehon (2010) puts it:

[21] I refer to this as the "consequential critique" of sceptical theism in [reference omitted]

> "*If the theist takes seriously the claim that God has good reasons for allowing so much suffering, then the theist should be the victim of moral paralysis: she should have no confidence in her moral judgments; she should have no idea when to allow suffering and when not to allow it, and she should also be unwilling to make moral judgments concerning the actions of others."* (emphasis original)

If this is right, then sceptical theism really does have significant practical and epistemic costs. Anyone who is committed to it is bound to confront ongoing, persistent moral paralysis. Of course, there is considerable debate about whether this is in fact right. Defenders of sceptical theism have tried to avoid the alleged costs (Anderson 2012; Bergmann & Rea 2005) of their position, and there are replies to those defenders saying that this is impossible (Maitzen 2013). I pass no judgment on who succeeds in this debate. My concerns, once again, lie with the potential parallels with the debate about AI doomsday scenarios.

In this respect, I want to highlight some of the structural aspects of the moral paralysis critique of sceptical theism. The most obvious structure feature of the critique is the chain of inferences it traces out from the core commitments of the sceptical theist. It starts by noting how the sceptical theist is committed to both the existence of God and to scepticism about the kinds of suffering it might be permissible for God to allow. It then draws out the implications of those commitments, by pairing them up with other beliefs about God. So, typically, the theist believes that God has some sort of creative and supervisory control over events unfolding in the world around us. This means that God could be intervening to prevent evil if He felt it was necessary. The fact that He is not, and that we don't know what his moral goals ultimately are, leads to our moral paralysis. It could be that God wants us to intervene, or wants things to continue as they are. We just don't know. This process of connecting sceptical theism to other beliefs and commitments, and then drawing out the implications of those combinations of beliefs and commitments, gives rise to the critique.

The second structural feature of the critique relates to what its proponents are trying to achieve. Obviously, none of them think that we should be morally paralysed or subject to some other undesirable consequence. What they are hoping is that their critique highlights the absurdity of the original set of beliefs and commitments. They

are trying to make sceptical theists abandon their response to the problem of evil by highlighting how deeply it conflicts with other cherished beliefs and commitments.

The question I wish to consider is whether these structural aspects of the critique have analogues in the case of AI risk.

4. What are the practical and epistemic costs of the treacherous turn?

In order to answer that question, I must confront some obvious disanalogies between the commitments of sceptical theists and the commitments of AI doomsday proponents. One reason why the commitments of sceptical theists are such fertile ground for arguments of the sort just outlined is that theism is itself a deep and pervasive commitment. Theism is a belief about the ultimate nature and cause of reality.[22] Consequently being a theist has implications, explicit or implicit, for nearly every other belief one holds[23] about that reality.

It's pretty clear that the commitments of AI doomsday proponents do not have similarly wide-ranging implications. A superintelligence would be a powerful being, to be sure, but believing in the possibility of its existence doesn't affect how one understands every other aspect of reality. If there are practical and epistemic costs associated with the treacherous turn, then they will be limited to how we interact with and understand AIs. But that limitation is not necessarily fatal to the argument I wish to make. Even if the consequences are limited to that domain, they could nevertheless be significant.

We can see this by considering, more deeply, the practical implications of the treacherous turn. Recall that the idea is that a superintelligent AI, with human unfriendly goals, could deceive us into thinking it poses no risk. It could do this in a number of nefarious ways. The most straightforward would be to "play nice", *i.e.* demonstrate its full capabilities but do so in a way that is wholly compatible with human flourishing. The slightly more duplicitous way would be to "play dumb", *i.e.*

[22] Schellenberg (2007) refers to beliefs of this sort as being forms of "ultimism"
[23] The one exception here might be beliefs about logical or mathematical truths, though there are theists who claim that those truths are dependent on God as well.

conceal its true intelligence and capabilities and trick us into thinking, even if its goals aren't wholly consistent with ours, it could never really pose a threat. It is this possibility that has, perhaps, the most significant practical and epistemic implications.

Look at what Bostrom has to say:

> "*[A]n unfriendly AI may become smart enough to realize that it is better off concealing some of its capability gains. It may underreport on its progress and deliberately flunk some of the harder tests, in order to avoid causing alarm before it has grown strong enough to attain a decisive strategic advantage. The programmers may try to guard against this possibility by secretly monitoring the AI's source code and the internal workings of its mind; but a smart-enough AI would realize that it might be under surveillance and adjust its thinking accordingly."*
> (Bostrom 2014, 117)

The only chink of light in all this is the fact that there may be a brief moment — Bostrom calls it the moment of the *conception of deception* (Bostrom 2014, 282) — before the AI realises that such wholesale deception is required. But even that chink can be quickly shut down:

> *"[H]aving had this realization, the AI might move swiftly to hide the fact that the realization has occurred, while setting up some covert internal dynamic (perhaps disguised as some innocuous process that blends in with all the other complicated processes taking place in its mind) that will enable it to continue to plan its long-term strategy in privacy*."
>
> (Bostrom 2014, 282)

Think about what this really implies. It implies that all our interactions with artificial intelligences should be shrouded by a deep and, arguably, paralysing suspicion. For it could be that an AI will, in the very near future, develop the level of intelligence necessary to undertake such a deceptive project and thereby pose a huge existential risk to our futures. In fact, it is even worse than that. If the AI could deliberately and conceal its true intelligence from us, in the manner envisaged by Bostrom, it could be that there

already is an AI in existence that poses such a risk. Perhaps Google's self-driving car, or IBM's Watson have already achieved this level of intelligence and are merely biding their time until we release them onto our highways or into our hospitals and quiz show recording studios? After all, it is not as if we have been on the lookout for the moment of the conception of deception (whatever that may look like). If AI risk is something we should take seriously, and if Bostrom's notion of the treacherous turn is something we should also take seriously, then this would seem to be one of its implications. That looks like a pretty significant practical and epistemic consequence to me.

What I want to suggest now is that this gives rise to two further arguments, each tending in opposite directions. The first is a *reductio*, suggesting that by introducing the treacherous turn, Bostrom reveals the underlying absurdity of his position. The second is an *a fortiori*, suggesting that the way in which Bostrom thinks about the treacherous turn may be the right way to think about superintelligence, and may consequently provide further reason to be extremely cautious about the development of artificial intelligence.

Let's start with the reductio. This argument is modelled on the critiques of sceptical theism I outlined in the previous section. Those arguments work by trying to get people to give up sceptical theism, but not necessarily by trying to get them to give up any other beliefs and commitments they might have. The reductio of Bostrom's argument works in the same way. It doesn't discount the seriousness of AI risk itself, it just discounts one particular set of views about the nature of that risk. It suggests that Bostrom's concept of the treacherous turn forces us to take seriously an absurd set of possibilities, and for that reason it should be discounted or revised. This has the further implication that the credibility of the empirical testing objection should not be impugned simply by reference to that concept. The argument works in the following way:

(1) If an argument or idea commits us to absurd beliefs, then it ought to be abandoned or revised.

(2) Bostrom's concept of the treacherous turn commits us to absurd beliefs.

(3) Therefore, Bostrom's concept of the treacherous turn ought to be abandoned or revised.

I take it that the first premise is relatively uncontroversial. The key to the argument is the second premise. What can be said in its favour? I have two suggestions. First, Bostrom's concept is absurd because it forces us to re-evaluate the kinds of inductive inference we routinely — and successfully — rely upon in our relationships with current (and, no doubt, near future) technologies.[24] This was the point I just tried to make by outlining, more fully, Bostrom's thoughts on what it would mean for an AI to take the treacherous turn by playing dumb. Now, of course, Bostrom wants us to re-evaluate those inductive inferences, but it seems absurd to do so since they are such a pervasive and, thus far, reasonably reliable feature of our AI-engineering projects. Second, and perhaps more convincingly, Bostrom's concept is absurd because it is effectively a restatement of classic Humean empirical scepticism, applied more restrictively to our interactions with AIs. Hume famously argued that we could never infer that the sun would rise tomorrow merely from the fact that it rose on all previous days. Bostrom is now arguing that we can never infer the future benevolence or safety of an AI merely from the fact that is has been safe on all previous occasions. There is, of course, a sense in which Hume is correct, and we can all appreciate the logicality of his argument, we routinely (and rightly) ignore it in our everyday lives because it would have absurd practical and epistemic consequences. Maybe the same is true of Bostrom's concept of the treacherous turn. What this highlights, I submit, is an absurdly low modal standard that operates in the debate about AI risk. Outlandish and merely possible scenarios take on a huge practical and epistemic significance given the beliefs about the capacities of a superintelligence. It is a serious question as to whether such a low standard is appropriate.

If premise (2) is acceptable, and the *reductio* argument as a whole works, it is still worth exploring the implications of the conclusion. As I said above, it doesn't imply that AI risk is irrelevant, and that we shouldn't worry about the prospect of deceptive AIs. It simply implies that we shouldn't worry about those things in the same way. In particular, we shouldn't worry about the kinds of outlandish hypothetical scenarios he

[24] Note how the focus here is limited to how the treacherous turn affects inductive inferences we make about artificial intelligences only. It does not affect all inductive inferences. This is unlike the situation with respect to sceptical theism.

entertains. We should feel more confident in continuing with the kinds of cautious, empirical testing that have stood us in good stead in the past. In this respect, the *reductio* argument chimes with other extant critiques of AI doomsayers. For instance, AI doomsayers sometimes worry that giving a superintelligent AI a goal like "maximise human pleasure" would lead it to hook us all up to a dopamine drip (or something similar). But others have complained that worrying about such possibilities is absurd because it misunderstands what it would take for something to count as superintelligent.[25]

Although the *reductio* argument has a certain appeal — not least in its ability to quell our existential fears — it is threatened by the *a fortiori*. This argument is slightly more difficult to outline. It starts by challenging the notion — central to the *reductio* — that Bostrom is wrong to entertain outlandish hypothetical scenarios in his effort to get us to take AI risks seriously. If you think about it, Bostrom's case for the treacherous turn, and the various forms it could take, depends heavily on a form of modal reasoning (reasoning about possible worlds). Bostrom argues that in order to properly plan for the threat of AI risk, we need to consider how a superintelligent AI might act across a range of possible worlds. Possible in what sense? Well, in a physical or technological sense, *i.e.* what would it be possible for a being with such powers to do? The scenarios Bostrom envisages might seem outlandish or absurd from our present perspective, but that perspective is one that deals with a very narrow range of possible worlds, viz. worlds that are possible for beings like us to create. We really have no idea how a superintelligence might reason, or the probability with which it might explore the kinds of possible world Bostrom is imagining. Kevin Warwick puts the point rather pithily:

> "*We won't really be able to understand why a superintelligent machine is making the decisions it is making. How can you reason, how can you bargain, how can you understand how that machine is thinking when its thinking in dimensions you can't conceive of?"*
>
> (Quoted in Barat 2013, 78)

[25] Richard Loosemore has made these complaints. I'm not aware of any academic publications in which he has made them, but he has done so in two online articles (Loosemore 2012 & 2014)

Indeed. And this is precisely the point that sceptical theists were originally making in their response to the problem of evil. The possible worlds of which we are aware are just a mere sliver (and perhaps not even a representative sliver) of all the possible worlds there are.

If Bostrom is right to imagine the kinds of outlandish possible worlds that he imagines, where does that leave us? It leaves us with the *a fortiori* argument. For if he is right, we should be far more worried about the prospect of superintelligence than even he seems to be. We should really be shutting down all existing artificial intelligence research and development projects, and crossing our fingers in the hope that one of our creations hasn't already crossed the threshold and made the treacherous turn.

Furthermore, this is one of the areas in which the disanalogies between sceptical theism and the treacherous turn may actually support a more radically sceptical stance in relation to superintelligences than in relation to God.[26] Critics of sceptical theism sometimes point out that the sceptical stance toward inductive inferences from our experience of the world around us are particular odd in light of God's supposed nature. God is supposedly the creator and sustainer of everything (or nearly everything) in existence. So everything we see, hear and experience in the world provides some evidence as to his true intentions. We can be sceptical of those inferences to some extent, but to completely undercut them – as sceptical theists seemingly do – would be absurd. But this is not true in the case of a superintelligence. A superintelligent AI would not be the creator and sustainer of everything in existence (though it may acquire great practical power). Consequently, not everything that we see, hear or experience would provide grounds for making inductive inferences about its ultimate intentions. Our observations of it in the contrived "box"-like environment provide the basis for some inductive inferences, but not for ones that are as robust as they would be in the case of God. Thus, this is a reason for thinking that the *a fortiori* argument is more persuasive in the case of AI risk and treacherous turn, than the *reductio*.

There will, of course, be responses to this. Even those who want us to take AI risk seriously will argue that some of this goes too far. They will argue that there are great

[26] I am indebted to an anonymous reviewer for encouraging me to make this point.

risks associated with superintelligence, but there are also great benefits.[27] We shouldn't abandon the project of developing AI simple because of these risks. But this doesn't seem right, not if we have to take the treacherous turn and its implications seriously. For if that is right, we may have no way of recognising whether the seeming benefits are actual benefits at all. Of course, you may argue that we will know them by simply experiencing them: if they seem beneficial to us then they are (because the gap between the seeming and the actual doesn't arise in the case of benefits accruing to us as individuals). And this would be correct, but only in the short term. If we need to be worrying about long term existential risks -- as AI doomsayers encourage us to be -- then we have no way of knowing for sure that an AI that confers benefits to us in the short-term is actually a net positive. That might be the most serious implication of the treacherous turn.

5. Conclusion

In conclusion, in this paper I have made three arguments about AI doomsaying (as practiced by Bostrom) and its implications. First, I argued that the epistemic move that Bostrom makes whilst defending the seriousness of AI risk is analogous to the epistemic move made by sceptical theists when responding to the problem of evil. In both instances there is an attempt to block the inference from seeming benevolence/evil to actual benevolence/evil. I then argued that blocking such an inference could have significant practical and epistemic implications. It may either imply that the kind of position taken up by Bostrom is absurd, and hence should be abandoned or significantly revised; or it could imply that the kinds of hypotheticals entertained by Bostrom should be taken far more seriously than they are. In other words, it either reduces Bostrom's position to the absurd, or strengthens it significantly.

I pass no judgment here on which of the two arguments is correct. Both have their appeal. The *reductio* appeals to our common sense and provides some reassurance about existential risk. The *a fortiori* makes some plausible claims about the type of modal reasoning that is appropriate in this debate. The key question is whether those claims really are plausible: what should the modal standard be? That is something participants in this debate need to address.

[27] This is the view of the *Machine Intelligence Research Institute* and some of its affiliated scholars, e.g. see Muehlhauser, L and Salamon, A. (2012).

Acknowledgements: The author would like to thank Stephen Maitzen and Felipe Leon for conversations about sceptical theism, and Alexander Kruel and an anonymous reviewer for feedback on a previous draft of this paper.

References

Almeida, M., & Oppy, G. (2003). Sceptical theism and evidential arguments from evil. Australasian Journal of Philosophy. 81: 496–516

Anderson, D. (2012). Skeptical theism and value judgments. International Journal for the Philosophy of Religion 72: 27–39

Armstrong, S. (2013) General Purpose Intelligence: Arguing the Orthogonality Thesis. Analysis and Metaphysics 12: 68-84

Barrat, J. (2013). Our Final Invention: Artificial Intelligence and the End of the Human Era. New York: St. Martin's Press.

Bergmann, M. (2001). Skeptical Theism and Rowe's new evidential argument from evil. Nous 35: 228

Bergmann, M. (2009). Skeptical Theism and the Problem of Evil. In T. P. Thomas & M. Rea (Eds.), The Oxford Handbook of Philosophical Theology. Oxford: OUP.

Bostrom, N. (2014) Superintelligence: Paths, Dangers, Strategies. Oxford: OUP

Bostrom, N. (2013) Existential Risk Prevention as a Global Priority. Global Policy 4: 15-31

Bostrom, N. (2012). The Superintelligent Will:Motivation and Instrumental Rationality in Advanced Artificial Agents. Minds and Machines 22(2): 71-85

Bostrom, N and Cirkovic, M. (eds) (2008) Global Catastrophic Risks. Oxford: OUP.

Bringsjord, S., Bringsjord, A. and Bello, A. (2012). Belief in the Singularity is Fideistic. In Eden, A., Moor, J., Soraker, J. and Steinhardt, E. (eds) Singularity Hypotheses: A Scientific and Philosophical Assessment. Dordrecht: Springer.

Danaher, J. (2014). Skeptical Theism and Divine Permission: A Reply to Anderson. *International Journal for Philosophy of Religion* 75(2): 101-118

Doctorow, C. and Stross, C. (2012) The Rapture of the Nerds. New York: Tor Books.

Dougherty, T. (2012). Recent work on the problem of evil. (2012) 71 Analysis 560-573; and

Dougherty and McBrayer (eds) (2014). Skeptical Theism: New Essays. Oxford: OUP.

Eden, A., Moor, J., Soraker, J. and Steinhardt, E. (eds) (2012). Singularity Hypotheses: A Scientific and Philosophical Assessment (Dordrecht: Springer 2012)

Hasker, W. (2010). All too skeptical theism. International Journal for Philosophy of Religion 68: 15-29

Loosemore, R. (2014). The Maverick Nanny with a Dopamine Drip: Debunking Fallacies in the Theory of AI Motivation. IEET, 24 July 2014 - available at: http://ieet.org/index.php/IEET/more/loosemore20140724 (accessed 31/10/2014).

Loosemore, R. (2012). The Fallacy of Dumb Superintelligence. IEET, 28 November 2012 - available at: http://ieet.org/index.php/IEET/more/loosemore20121128 (accessed 31/10/2014).

Lovering, R. (2009). On What God would Do" (2009) International Journal for the Philosophy of Religion 66(2): 87-104

Maitzen, S. (2013). The Moral Skepticism Objection to Skeptical Theism. In McBrayer, J. and Howard-Snyder, D. (eds) A Companion to the Problem of Evil. Oxford: Wiley-Blackwell.

McBrayer, J (2010). Skeptical Theism. Philosophy Compass 5: 611-623

Muehlhauser, L and Salamon, A. (2012). Intelligence Explosion: Evidence and Import. In Eden, A., Moor, J., Soraker, J. and Steinhardt, E. (eds) (2012). Singularity Hypotheses: A Scientific and Philosophical Assessment (Dordrecht: Springer 2012)

Piper, M. (2008). Why theists cannot accept skeptical theism. Sophia 47(2): 129-148

Rowe, W. (1979). The Problem of Evil and Some Varieties of Atheism. American Philosophical Quarterly 16(4): 335-341.

Schellenberg, J.L. (2007) The Wisdom to Doubt. Ithaca, NY: Cornell University Press.

Sehon, S. (2010). The problem of evil: Skeptical theism leads to moral paralysis. International Journal for the Philosophy of Religion 67: 67–80.

Street, S. (forthcoming). If there's a reason for everything then we don't know what reasons are: Why the price of theism is normative skepticism. In Bergmann and Kain (eds) Challenges to Religious and Moral Belief: Disagreement and Evolution. Oxford: OUP.

Trakakis, N. (2007) The God Beyond Belief: In Defence of William Rowe's Argument from Evil. Dordrecht: Springer.

Wielenberg, E. (2010). Sceptical Theism and Divine Lies. Religious Studies 46: 509-523

Wielenberg, E. (2014). Divine Deception. In Dougherty and McBrayer (eds) (2014). Skeptical Theism: New Essays. Oxford: OUP

Wykstra, S. (1996). Rowe's Noseeum Arguments from Evil. Reprinted in Howard-Snyder, D. (ed) The Evidential Argument from Evil. Bloomington, IN: Indiana University Press.

Yampolskiy, R. (2012) Leakproofing the Singularity. Journal of Consciousness Studies 19: 194-214

Yudkowsky, E. (2008) Artificial Intelligence as a Positive and Negative Factor in Global Risk. In Bostrom, N and Cirkovic, M. (eds) (2008) Global Catastrophic Risks. Oxford: OUP.