Journal of EVOLUTION & Technology

A peer-reviewed electronic
journal published by the Institute
for Ethics and
Emerging Technologies
ISSN 1541-0099
27(1) – June 2017

Superintelligent AI and Skepticism

Joseph Corabi
Department of Philosophy
Saint Joseph's University

jcorabi@sju.edu

Journal of Evolution and Technology - Vol. 27 Issue 1 – June 2017 - pgs 4-23

Abstract

It has become fashionable to worry about the development of superintelligent AI that results in the destruction of humanity. This worry is not without merit, but it may be overstated. This paper explores some previously undiscussed reasons to be optimistic that, even if superintelligent AI does arise, it will not destroy us. These have to do with the possibility that a superintelligent AI will become mired in skeptical worries that its superintelligence cannot help it to solve. I argue that superintelligent AIs may lack the psychological idiosyncracies that allow humans to act in the face of skeptical problems, and so as a result they may become paralyzed in the face of these problems in a way that humans are not.

1. Introduction

Among philosophers, no one has given more attention to the dangers and opportunities of the "intelligence explosion" than Nick Bostrom. In his recent work *Superintelligence: Paths, Dangers, Strategies*, Bostrom pursues one of the first book-length philosophical investigations of the looming prospect of Superintelligent Artificial Intelligences (hereafter SAIs) – non-biological machines with systematically greater cognitive skill than humans.[1] He believes that SAIs are likely coming in decades (or at most centuries) rather than in millennia or never. And while the picture he paints is not all doom and gloom, he is often quite pessimistic about what SAIs would mean for human civilization, and even human life itself. Because SAIs would be systematically smarter than human beings, it would be relatively easy for a malevolent SAI to manipulate us, gain control of resources, and wreak widespread havoc on the world. As we will see, it is also very hard to guarantee that an SAI would be benevolent enough to avoid a terrible outcome. This is because SAIs, in virtue of the enormous power and influence they would be likely to acquire, could very easily destroy human civilization and cause the extinction of humanity even without making human destruction a high priority. Such an apocalyptic scenario could arise merely as a result of an SAI's ruthless efficiency in pursuing some seemingly trivial and harmless goal.

The work of philosophers like Bostrom and David Chalmers (2012) has spurred increasing speculation about the likelihood of attaining SAI in the near future and debate about whether such a development would be positive, negative, or neutral for practical human interests. The tendency is to favor one of two diametrically opposed views – either that SAIs will be extraordinarily positive or extraordinarily negative for us (sometimes remaining agnostic about which while affirming that it will almost definitely be one of the two). My goal in this paper is to explore the third option – that SAIs might wind up being neutral for us. How might this occur? SAIs might wind up being *paralyzed* from acting due to problems or conflicts in their own motivational schemes. In the end, I will not try to demonstrate that SAIs *will be* paralyzed or even try to show that paralysis is especially likely, but I will argue that it is an underappreciated possibility.

In the event that an SAI does wind up acting (which might be either very positive or very negative for humanity), there are also lessons to be learned about exactly what sort of cognitive agent the SAI is. In particular, depending on the circumstances, the SAI's failure to be paralyzed could indicate that the SAI has cognitive weaknesses rather than cognitive strengths.

After a brief introduction to the problems posed by superintelligence, I explore the central issue: potential failures of the SAI to escape from skeptical problems.

2. The problems posed by Superintelligent Artificial Intelligence: A brief introduction

We can begin to appreciate the dangers and opportunities presented by SAIs by noticing that, all else being equal, great cognitive skill in pursuit of most goals leads to a higher likelihood of success than normal cognitive skill does. A highly cognitively skilled agent pursuing a goal will gather more and better information than a normal agent, as well as developing superior strategies. Insofar as such an agent competes with lesser cognitive agents, it will also be able to locate strategic advantages in order to manipulate them into making mistakes. It will thus be able to amass more and more power and influence. If the environment presents cognitive obstacles that are challenging enough and the agent's cognitive advantage is great enough, it might even be able to single-handedly go from a modest starting point to domination of its environment. (For much more detailed treatment of the basic ideas in this section, see Chalmers 2012 and Bostrom 2014.)[2]

Consider, for example, an AI that was vastly cognitively superior to any human being, and which had the goal of dominating Earth. Such an agent could begin its "life" as an isolated piece of hardware. It could then trick a human into giving it access to the internet, whereupon it could start amassing information about economics and financial markets. It could exploit small security flaws to steal modest amounts of initial capital or convince someone just to give it the capital. Then it could go about expanding that capital through shrewd investment. It could obtain such an advantage over humans that it might then acquire so many resources that it could start influencing the political process. It could develop a brilliant PR machine and cultivate powerful connections. It might then begin more radical kinds of theft or even indiscriminate killing of humans that stood in its way of world financial domination, all the while anticipating human counter-maneuvers and developing plans to thwart them.[3] (How might it kill humans? It could hire assassins, for instance, and pay them electronically. Or it could invent and arrange for the manufacture of automated weapons that it could then deploy in pursuit of its aims.)

As I mentioned above, even seemingly innocuous or beneficent goals could result in similarly catastrophic results for human beings. Imagine, for instance, Bostrom's example of an SAI that has the seemingly trivial and harmless goal of making as many paper clips as possible (see Bostrom 2014, 107–108 and elsewhere). Such an SAI might use the sorts of tactics described above in a ruthless attempt to amass resources so that paper clip manufacturing could be maximized. This sort of agent might notice that human bodies contain materials that are useful in the manufacture of paper clips and so kill many or all humans in an effort to harvest these materials. Or, it might simply see humans as a minor inconvenience in the process of producing paper clips, and so eliminate them just to get them out of the way. (Perhaps humans sometimes obstruct vehicles that deliver materials to automated paper clip factories, or consume resources that could be used to fuel these factories.) Even an SAI that had the maximization of human happiness as an ultimate goal might easily decide the best course of action would be to capture all the humans and keep them permanently attached to machines that pumped happiness-producing drugs into them. The SAI would then go about managing the business of the world and ensuring that the support system for the human race's lifestyle of leisure did not erode.

Obviously, the flip side of the worries just discussed is that an SAI that aided humans in the pursuit of their genuine interests could be a powerful positive force – it could cure diseases, even save humans from impending existential risks, and make human lives happier and more fulfilling.

The issue of controlling an SAI and ensuring that it did not pursue goals that would likely result in catastrophe for human beings is known as the "control problem" (Bostrom 2014). Ideally, designers of AI would identify a pre-set goal that lines up with human interests and then directly program an AI to have it. This is the Direct Approach, which we can roughly define as follows:[4]

> The Direct Approach: The designers of the AI engineer the AI to have an ultimate goal that involves no appeal to any desires or beliefs of any persons or groups to specify the content of that goal. (An example would be programming the AI to have the goal of maximizing human happiness.)

Unfortunately, it turns out to be very hard to do this in a way that doesn't lead to disaster. This is for reasons closely related to problems philosophers face concerning the difficulty of conceptual analysis – even after one has what appears to be a sharp grasp of some concept or intuitive phenomenon, it proves to be extraordinarily difficult to spell out a set of conditions that are perfectly coextensive with the concept or intuitive phenomenon supposedly grasped. (See, for instance, the remarks in Russell 1918. For a discussion of related points, see Bostrom 2014, 120.) And because of the potential ease with which SAIs might amass power, when we give our explicit instructions to the SAI, any seemingly small failures to capture an intuitive idea we have about what its final goal should be might result in huge divergences from the sort of behavior we envisioned the SAI engaging in.[5] (It is not hard to imagine an AI engineer thinking that an SAI programmed to maximize human happiness would create something that everyone would recognize as a paradise, for example. But on reflection we see that things might easily go very differently.)

In an attempt to solve the control problem, some theorists have proposed dispensing with direct specifications of an SAI's ultimate goals in favor of indirect specifications. These can be understood (again roughly) as follows:

> The Indirect Approach: The designers of the AI engineer the AI to have as its ultimate goal to achieve what some group wants or believes to be best (or would want or believe to be best under specific idealized conditions) or what the AI itself believes to be what the truth about morality demands.

Bostrom, for instance, advocates the indirect approach for solving the control problem and sympathetically discusses several indirect options. One is the possibility that we might program an SAI to have as a final goal to promote whatever humanity as a whole would want to promote under idealized conditions. Another option is the possibility that we might program it simply to do what is objectively morally right or morally best, then rely on its superior cognitive skill to discover what the morally right or morally best thing is. (See the discussion in Bostrom 2014, 212–20.)

Throughout the rest of the paper I will investigate problems that apply primarily to these indirect approaches – specifically, to the claim that an SAI that follows one of these approaches will succeed in acting. (Much of our investigation is also relevant to direct approaches, but since they are out of favor – justifiably in my view – I won't spend much time addressing them.) To aid our discussion, I will distinguish between indirect approaches that proceed by trying to discover what some group of agents (most likely humans) wants or thinks (either in reality or in idealized circumstances) and indirect approaches that proceed by trying to discover truths about morality. Call the former "Indirect Psychological" (IP) approaches and the latter "Indirect Moral" (IM) approaches.[6] The problem we will examine applies to either approach, but the nuances differ somewhat depending on which is our focus.

The problem of skeptical superintelligences

Our problem could occur in tandem with any kind of indirect approach. In fact, it could easily affect any AI employing a direct approach as well, but since direct approaches are generally not favored, I will not spend much time concentrating on them.

So let us turn to the problem of skeptical SAIs. A skeptical problem can be thought of as a challenge to the truth or rational plausibility of a claim (or, more generally, to the truth or rational plausibility of a type of claim – potentially quite broad).[7] Typically, these challenges are thought to require appeals to further evidence or other rational considerations on the part of the agent considering the claim or type of claim.

The potential role of skeptical problems in the psychology of an AI – particularly a superintelligent AI – has been underappreciated. Consider the following argument:

> (1) There are some skeptical problems that humans haven't solved.
>
> (2) There is no compelling reason to think that superintelligence is going to be much more useful than regular human intelligence in solving skeptical problems.
>
> (3) If (1) and (2), then it is likely that SAIs will fail to solve the skeptical problems that humans have failed to solve.
>
> So, from (1)–(3):
>
> (4) It is likely that SAIs will fail to solve the skeptical problems that humans have failed to solve.
>
> (5) Humans have psychological idiosyncrasies that allow them to act in the face of unresolved skeptical problems.
>
> (6) SAIs might easily lack those psychological idiosyncrasies.
>
> (7) If (4)–(6), then it is at least fairly likely that SAIs will be paralyzed – i.e., unable to take concrete actions (actions aside from thinking, insofar as thinking counts as an

action). (There might also be some initial period where the AI is able to take actions that involve information gathering, but this is likely to require the introduction of significant subtlety to the analysis. The issues raised are important, but would take us too far afield.)

So, from (4)–(7):

(8) It is at least fairly likely that SAIs will be paralyzed.

While I do not claim that this argument is sound, I also do not think there is any obvious reason to believe it is unsound (in the sense of a reason that should readily occur to competent professional philosophers or AI researchers, or which has appeared in some uncontroversial form in the philosophical or AI literature). My aim is to encourage further reflection on it, in the hope that philosophers and AI researchers can arrive at a confident and informed judgment about its plausibility.

While I cannot address all of the vast array of subtleties – both empirical and philosophical – that must be tackled to offer an ultimate assessment of the argument, it is worth briefly considering some of the premises individually to appreciate why they appear worthy of serious consideration. (I will wait until the next section to consider two of the key premises – (5) and (6) – owing to the complexity of evaluating them.)

Premise (1) clearly seems plausible. Although there may be some controversy over the ultimate fate of this or that skeptical problem, there is little reason for thinking that every important skeptical problem in philosophy has been solved – no matter how one counts, there is a multitude of these problems, and virtually all have had serious and persistent philosophical defenders. Here are a few representative examples that have at least a reasonably good claim to be unsolved:[8]

(A) *The Problem of Moral Properties*. Moral properties do not appear identifiable with any kind of non-normative property. They seem, in other words, to be the wrong sort of entity to be identifiable with any such property. Nor do they appear to stand in any other relation to non-normative properties (such as being realized by them) that might make the normative properties metaphysically unproblematic or give us a way of gaining epistemic access to or information about them (via sensory evidence). How then can we come to have justification for believing that these moral properties exist – and by extension for believing that any substantive moral claims are true, since *ex hypothesi* these claims are made true by the moral properties? (See Mackie 1977.)

(B) *The Problem of Justifying Beliefs About Sensations (The Sellarsian Dilemma)*. There are some reasons for thinking that the only way to offer a rigorous theory of the empirical world depends on reconstructions that begin from beliefs directly about one's own sensations, since this is the only foundation for empirical beliefs that one has in one's awareness. But when one considers the relationship between one's beliefs about one's sensations and the sensations themselves, a dilemma arises. Either the sensations that allegedly form the ultimate justificatory basis for one's beliefs about them are representational – i.e., the sort of thing that admits of truth or falsity – or not. If they are not representational, then how could they serve as a justificatory basis for the beliefs, since a justificatory basis is by its nature a *reason* to believe something? (And reasons are by their very nature allegedly the sorts of things that admit of truth or falsity.) But if, on the other hand, the sensations are representational, they can serve as a justificatory basis for the beliefs, but then they are themselves in need of further justification. Nothing further seems available to justify them, however.[9] (The Sellarsian dilemma was first

> described in Sellars 1956 and later discussed at length in Bonjour 1985 and Devries and Triplett 2000.)
>
> (C) *The Problem of Induction (Combined with The New Riddle of Induction)*. The Problem of Induction arises anytime one wishes to generalize from some observed sample to a wider population.[10] In order to infer that the unobserved "region"[11] will be like the observed "region" in the relevant respects (or more generally, that the unobserved "region" is more likely to have some particular characteristics rather than others),[12] grounds are required to justify the assumption that some particular projection or hypothesis is to be privileged as more epistemically likely than others. Some general feature is then advanced in an attempt to justify this assumption. Most often this feature is simplicity, but there are other candidates as well. However, typically the only candidate for rigorous defense of the claim that this feature is an indication of likelihood of truth is to generalize from its success in other attempted projections. (E.g., in the past, when we were up in the air about two hypotheses at time *t* but then got more evidence at *t+1*, more often than not the simpler one wound up being true, or at the very least it was the one that wound up receiving more support from the new evidence.) This seems viciously circular, because it begins with an attempt to justify generalizations or projections in general, a justification is given for the preferred method of generalization or projection, and then this justification is in turn justified by appealing to previous generalizations or projections.[13] (See Hume 1978 (1738) for the original discussion of the problem of induction, and see Goodman 1955 for the New Riddle of Induction.)

Admittedly, it is harder to offer a confident assessment of premise (2) above – that there is no compelling reason to think that superintelligence is going to be much more useful than regular intelligence in solving skeptical problems – because it is difficult to predict in advance how well suited to solving skeptical problems an SAI would be. If the problems are like a difficult puzzle – akin to a massive Rubik's Cube or convoluted chess problem – then it is plausible to suppose that the SAI will enjoy a significant advantage over human beings. Surely the massive computational power and memory possessed by the SAI would make it well-suited to solving puzzles of this sort, even ones that have seemed hopeless to humans. But it may be that what makes the skeptical problems difficult is not that they are massively computationally difficult puzzles, but that the agent is positioned so that a priori reasoning along with sensory data is not enough to make serious progress, no matter how impressive the reasoning or the data. That would make tremendous intelligence and information of the sort accessible to the SAI useless in solving them. Getting to the bottom of the usefulness of the SAI's superintelligence might be a complicated and difficult undertaking. I will set it aside for now, flagging it as an area for further reflection.

Premise (3) – the claim that if the previous two premises are true, it is likely that SAIs will fail to solve the outstanding skeptical problems – is fairly straightforward. Since humans seem far from offering solutions to these problems, even a small advantage on the SAI's part would not much improve its chances of arriving at a solution.

Premise (7) is similarly straightforward, at least in outline. If SAIs are likely to fail in their efforts to solve skeptical problems (just like humans), but humans have psychological idiosyncrasies that allow them to act in spite of their epistemological failures, it would stand to reason that, if SAIs lacked those idiosyncrasies, there is a good chance that they would be unable to act. There remains one issue that needs to be addressed to motivate (7), however – namely, the issue of why these skeptical worries would cause trouble for the SAI to begin with. Why, in other words, would a failure to solve the skeptical problems potentially lead an SAI to paralysis, presupposing it did not have some psychological idiosyncrasy that would allow it to ignore its conundrum and act anyway?

As with any agent, if the SAI is to perform actions it will require some ultimate goal that is directing it. As I alluded to before, SAIs will have ultimate goals that are pre-set. The outline of these goals will be specified either directly or indirectly by the SAI's programmers. Recall that we are examining indirect approaches here, where the SAI's programmers specify that the SAI should have as its ultimate goal whatever humans would most want or believe is best, perhaps under various idealized conditions (an IP approach), or whatever is most in accord with the true moral theory it discovers (an IM approach).

The specific answer to our question about why skepticism causes a problem for SAI action then depends on what the SAI's pre-set ultimate goal is. If the SAI is following an IP approach, then inevitably its pre-set goal will require it to answer a variety of empirical questions in order to discern what concrete actions to take. These would include questions about what the relevant group wanted or believed, and also questions involved in instrumentally extending fundamental desires associated with the pre-set goal. For instance, if the pre-set goal involved doing what humans would want on intrinsic grounds under various idealized conditions, the SAI would have to figure out what that thing was. Suppose (to use a toy example) that it were to turn out that what they would want under those conditions is to maximize net expected human pleasure in the long run. In that case, the SAI would also have to discern what would maximize net expected human pleasure in the long run. This would of course require a great deal of knowledge in psychology, biology, chemistry, physics, and a host of other disciplines.

In its efforts here, the SAI could be tripped up by skeptical problems at multiple junctures. First, if the SAI is not able to solve the problem of induction (combined with the New Riddle of Induction) described above, then the SAI could easily become mired in skepticism about the external world and worried that it was located in a mere simulation. It could also suffer from other forms of skeptical doubt, such as about whether humans have minds (and if so, what mental states humans are in). The reason for this is that the SAI could easily face a situation where all of its evidence is compatible with infinitely many hypotheses, and lacking a way to ground an assignment of a priori epistemic probabilities in those hypotheses, it could fail to have any rigorous reason to judge any of them more likely than the others.[14] These problems would be exacerbated if the SAI also fell into the Sellarsian Dilemma and began to question even its beliefs about its own sensations or sensory data.

To state what is perhaps now obvious, an SAI that could not answer empirical questions of this sort would have great difficulty acting. An SAI following an IP could not come to know empirical propositions required to discern what the relevant group wants or believes, nor could it come to know the empirical information required to connect actions under its control with its fundamental goal, even if miraculously it did come to discern what its concrete ultimate goal was supposed to be. (To use the example above, even if the SAI knew that the required course of action in accord with its IP was to maximize net expected human pleasure, it could not answer the empirical questions it would need to in order to figure out what specific actions would accomplish that maximization.)

An SAI following an IM would be no better off – in fact, it would likely be in even worse shape. An SAI following an IM would not be able to figure out what specific actions satisfy the empirical requirements for rightness or permissibility that the true moral theory lays out.[15] Suppose (again, to use a toy example) that the true moral theory discovered by the SAI claimed that an action was right if and only if it maximized net expected human pleasure. An IM SAI that could not answer empirical questions about the likely effects of an action on human pleasure could not act, because it could not link its ultimate goal (i.e., to act rightly, in accord with the moral propositions whose truth it has discovered) with any specific actions that it was capable of taking.[16]

The skeptical worries might not be restricted to empirical questions. Various questions that might not involve any empirical component – such as, arguably, the issue of whether moral realism is true – could be subject to skeptical problems that the SAI would fail to solve. We have already seen the skeptical worry about justifying belief in abstract moral entities, for instance.[17]

Note also that the chain here is very fragile. It might take only one skeptical roadblock for the SAI to become paralyzed. For instance, an IM SAI could succeed in answering all of the empirical varieties of skepticism, only to crumble because it fails to discover that moral realism is true or fails to come to know any specific moral truths.

Note also that the worries here do not affect only SAIs pursuing indirect strategies, lest one suppose that we have hit upon a reason to prefer direct approaches. Any specific ultimate goal that an SAI is directly programmed to have will still need to be instrumentally connected with actions that the SAI is capable of taking. For instance – to continue in the spirit of the examples above – if an SAI is directly programmed to have the ultimate goal of maximizing net expected human pleasure, it will need to answer various empirical questions about what that maximization would consist in and how to connect whatever that was with basic actions that it was capable of pursuing. This would require confronting the skeptical worries. (Again, all of our discussion here has presupposed that SAIs do not have psychological idiosyncrasies that allow them to act in the absence of rigorous justification for their beliefs. This is in keeping with the conditional nature of premise (7), a premise whose general plausibility I have been defending.)

I have now examined and explained why a variety of premises in the argument above should be accepted or at least taken seriously. Only premises (5) and (6) remain unexamined – the claims that humans have psychological idiosyncrasies that allow them to act in the face of unsolved skeptical problems, and that SAIs would be likely to lack those same idiosyncrasies. I will address them in the next section.

Psychological idiosyncrasies

If the previous discussion is on track, it is reasonable to conclude that some skeptical problems remain unsolved by human beings. Even if they have been solved, only a tiny segment of the human population possesses the solutions. Yet humans manage to ignore skeptical problems (or remain blissfully ignorant of their existence) and act anyway. This suggests – in keeping with premise (5) in the argument set out above – that humans have psychological idiosyncracies that facilitate this action. We must now address whether this is in fact the case, and if so, whether SAIs would be likely to replicate these idiosyncrasies. (Recall that premise (6) claims that they would not.)

Evaluating these issues involves getting to the heart of many of the most central issues surrounding the general topic of motivation and action in AIs. I do not offer any definitive pronouncements here. A first thing to notice, however, is that SAIs will be much more intelligent than human beings across the board, pretty much by definition. As just noted, in many cases where humans act in the face of skeptical problems that undermine the justification for beliefs (either normative or empirical) relevant to that action, the humans manage to act precisely because they are unaware of the skeptical worries. Sometimes this lack of awareness is more or less permanent, as with unreflective individuals, and sometimes it is temporary, as with philosophers who step out of their offices to play backgammon and dine with their friends (as per the famous remark in Hume's *Treatise*). Insofar as humans can manage to act as a result of such a lack of awareness, this is testament to a *deficiency* in their cognitive skills. The humans manage to act because they have failed to see (or are at least currently not seeing) a real epistemological problem. This seems like a genuine idiosyncrasy if anything does. Considering the greatly superior cognitive skills of an SAI, it seems implausible that an SAI would suffer from the same kind of deficiency, at least without being deliberately engineered in a way that gave rise to cognitive blindspots.[18] This issue will be discussed below. For the moment, though, suffice it to say that I have concerns both about the classification of such an agent as an SAI (a relatively superficial matter) and about its coherence and stability, given that it would have extraordinary cognitive skills in a wide variety of domains while having noteworthy cognitive shortcomings in a few specific areas (a much more substantive issue).

Of course, there are probably cases where humans manage to act in spite of being aware of the skeptical problems that plague them. In these cases, arguably humans are not displaying a cognitive deficiency, since arguably they are not endorsing incorrect judgments about their epistemic positions or making epistemically unjustified endorsements of propositions relevant to their action.[19] But humans in such scenarios appear to be displaying a peculiar form of unreasonableness – they seem to be acting *as though* they believed propositions that they do not really believe.[20] And perhaps they are even displaying a cognitive deficiency by outright believing (as opposed to merely acting as though they believed) propositions that they are aware are not made epistemically likely or certain by the available rational considerations. Either way, again we seem to be looking at a genuine idiosyncrasy.

Could an SAI manage a similar form of unreasonableness? This would largely depend on whether the SAI could (a) use great cognitive skill in evaluating the rational support for specific propositions and apportioning confidence based on that rational support, only to disregard that process when it came time to act, substituting some variety of pretend belief for real belief or else (b) sustain a genuine belief in a proposition that it was aware was not justified by rational considerations.

There is not much to commend a suggestion along the lines of (b), at least not in a way that is thoroughgoing enough to provide the SAI a path to action. To get to the bottom of the issue, we must deal separately with how the SAI would handle moral beliefs and empirical beliefs.[21] This leads us to subtly different conclusions for IP and IM SAIs, but the basic lesson is the same for both.

It is unclear whether irrationality at the level of fundamental moral beliefs would be sustainable for an SAI. (This would be of primary relevance for an IM SAI, since if it were mired in skeptical doubts about fundamental moral beliefs, it would be unable to generate any concrete ultimate goal from its pre-set ultimate goal. This is because it would be unable to discover any moral truths, and its pre-set goal would tell it to adopt as its concrete ultimate goal whatever it discovered was morally right or morally best.)[22] Much would depend on whether the products of this irrationality could be isolated from bringing about serious incoherence in other areas – areas that could potentially have relevance to broader cognitive processes that the SAI would need to maintain in order to preserve its superior overall intelligence and cognitive performance.

Whatever we ultimately conclude about whether this is feasible with moral propositions, when it comes to empirical propositions there is little reason to suppose that the SAI's intelligence would allow it to adopt genuine beliefs that were unreasonable, because there are more grounds to be confident that serious problems are on the horizon. This is of great significance, because it potentially affects both IP and IM SAIs in systematic ways. (If an IP SAI were constrained to be reasonable but remained mired in skeptical doubt, it would be prevented from discerning what concrete final goal it should have, because it could not answer empirical questions about what the relevant group wanted or believed – e.g., what humans would want under various idealized conditions. And all SAIs in this boat, whether IM or IP, would be prevented from connecting moral beliefs or ultimate desires with specific actions they are capable of taking.)

An SAI that did not apportion its confidence in empirical propositions based on rational considerations would probably become mired in a variety of incoherencies. These incoherencies would likely spell serious trouble for the SAI as it attempts to model the world accurately, given the tremendous sophistication and complexity that those attempts would have in comparison to the ones human beings make. Consequently, the SAI would likely not tolerate them and would take strong measures to prevent or eliminate them, judging an inaccurate worldview a major obstacle in achieving whatever ultimate goal it had.

Thus, given these issues, the best bet for the hunch that the SAI could manage to reproduce human unreasonablenesss would be some version of (a) – i.e., for it to have some worked-out method of adopting pretend beliefs for the purposes of action, without letting these in contaminate its pristine “real” beliefs. But the details of understanding the nature of this would require a great deal of subtlety. For present purposes, it will suffice to note that much further work would need to be done to render plausible the idea that an SAI could manifest the same kinds of irrationality in action as humans.

Someone might protest this general line of reasoning, though. After all, an SAI might be based originally on human biology – it could be produced by the computer emulation of an entire human brain, for instance, with subsequent enhanced versions made by human designers or the previous emulations themselves. (See, for instance, the discussion of whole brain emulation in Bostrom 2014, 30–36 and Chalmers 2012.) Wouldn’t this reproduce the characteristic unreasonableness – the distinctive psychological idiosyncrasy, whatever it is – that allows humans to act in the absence of good epistemological grounding? It might, but there is reason to be suspicious. As enhanced versions of artificial intelligence get made, the designers (be they humans or previous AIs) will need to make adjustments to the engineering and architecture in order to achieve increased cognitive skill. There is a good chance that these changes will be dramatic in the long run, as dramatic changes may well be needed to produce dramatic cognitive improvements. But as the process of rewiring progresses, there is an increasing chance that whatever architecture allows for the human idiosyncrasies will be lost.

Even setting the rewiring issue aside, we should notice that there is a commonly held stereotype (and modest empirical evidence) that smarter humans are more likely than less smart ones to “overthink” both decisions and everyday empirical questions, resulting in increased indecisiveness.[23] If there is even a shred of truth in this stereotype, what will happen when we extrapolate to the massive level of intelligence possessed by an SAI? (The SAI could, of course, benefit from greatly enhanced speed in resolving the everyday epistemological hangups that humans struggle with. But what if its massive intelligence gave rise to further hangups that its superior speed was not useful in resolving?)

I hope the discussion thus far has at least motivated the idea that the argument for paralysis should be taken seriously. In short, I hope that it has convinced readers that whether or not an SAI will succeed in acting may come down to whether it can solve a host of skeptical problems, some of which have vexed human philosophers for centuries, and some of which human philosophers have despaired of ever offering a rigorous solution to. I recognize, though, that there may be lingering doubts about the plausibility of a number of claims that I have entertained above. For that reason, let us consider some further objections:

> *Objection (A) – An SAI does not need to achieve certainty to act. As long as it is able to judge particular hypotheses or propositions more likely, it will have sufficient information to engage in decision-making under normal circumstances. And the sorts of skeptical problems you are discussing would only impede the SAI from achieving certainty, not from making judgments about likelihood.*
>
> Response – The skeptical problems at issue here are not skeptical problems that merely prevent an agent from arriving at absolute certainty. They prevent an agent from arriving at any degree of confidence in the relevant claims. The problem of induction, for instance, when combined with the New Riddle of Induction, does not merely prevent the agent from knowing that the conclusions of its inductive inferences will certainly hold. It prevents the agent from judging that the pattern represented by the conclusion of the inductive inference (e.g., "All swans are white," "10% of people have gene X") is more likely than any other pattern that is logically consistent with the data. The same goes for hypotheses designed to explain a set of evidence. Thus, if the SAI were to fail to solve this problem, it could not judge that the existence of an external world was more (or less) likely than its being deceived by a demon or its existing inside a complex simulation. This is because it would be unable to appeal to the simplicity of one of the global hypotheses over the others to ground a judgment that it is overall more likely. And I use simplicity just as an example here – the same would go for any other feature that might be a candidate for justifying the judgment that one hypothesis is more likely than another when both are equally supported by the total evidence. (The same applies to judgments about the hypotheses being equally likely, lest someone suppose that the SAI could simply default to being indifferent among the various hypotheses. Incidentally, this wouldn't help much anyway, because if there are uncountably many detailed hypotheses that are compatible with the data – and there surely are in most or all cases – this would not allow a coherent assignment of credences.) All the hypotheses would be consistent with the data. And without rational a priori grounds to prefer one hypothesis to another, when the evidence comes in it will be impossible to prefer any hypothesis a posteriori either.[24] This would plague not just the SAI's attempts to answer the most global of empirical questions – it would plague every attempt at inductive reasoning and inference to the best explanation that the SAI engaged in.
>
> *Objection (B) – In recent years, there has been increasing philosophical discussion of how to act under uncertainty. Philosophers like Andrew Sepielli have begun to work out theories of action under uncertainty that promise to offer rigorous guidelines for acting when one has some confidence in one course of action and some confidence in other courses of action. (See Sepielli 2009 and 2013, for instance.) Even if the SAI can't solve the skeptical problems, surely it will be able to use such a theory (or a successor to such a theory that it develops with its own superintelligence) and find a path to action that way. Thus, premise (7) is false because the SAI can fail to solve the skeptical problems, fail to replicate human psychological idiosyncrasies, but still manage to act.*
>
> Response – Things are not so simple. First, there may be distinctive skeptical problems that plague attempts to discover the correct principles for acting under uncertainty. Even setting this issue aside, theories like Sepielli's presuppose that agents have specific non-zero degrees of belief in particular propositions, or at least something that closely

approximates having such degrees of belief.[25] But many of the skeptical problems that we are discussing would prevent an SAI from achieving this modest goal. If they were to go unsolved, the most far-reaching of the skeptical problems (such as the Kripkenstein problem) would even prevent the SAI from pinning down the content of its own thoughts, let alone assigning justified credences to that content. Even setting aside these kinds of catastrophic skepticism, the SAI would be in the position classically described as "acting under ignorance" rather than "acting under risk." And most philosophers would agree that acting under ignorance is an area where little rigorous progress has been made. (See the basic discussion of acting under ignorance in Resnik 1987, chapter 2.) Acting under ignorance is acting when one has no justified level of confidence in a proposition, even a minimally precise one. Acting under risk is acting when one has a justified precise level of confidence in a proposition, but that level of confidence is less than 1 (e.g., .6). Also, traditionally "acting under risk" refers to risk or uncertainty on the empirical side, not the normative one. But it is natural to extend the meaning to normative risk or uncertainty as well.)

Objection (C) – You write of the SAI coming to gain knowledge of propositions and imply that this requires the SAI to solve all the various skeptical problems. But there is a robust tradition in the recent history of philosophy of supposing that individuals can come to know things without solving rigorous skeptical problems. For example, various forms of reliabilism entail that agents can have knowledge of many types of claims, provided that (roughly speaking) their beliefs were produced reliably. (See, for instance, Goldman 1976.) This is so even if the agents are not aware of any evidence or other reasons that allow them to refute skeptical challenges like the ones discussed above. Consequently, it is plausible to suppose that the SAI will be aware of these philosophical positions, persuaded by one of them, and go about acting without concerning itself with the skeptical problems (at least for the purposes of action).

Response – It is plausible that the SAI will be aware of these positions and the various defenses of them that can be constructed. It is less clear that the SAI will find them persuasive, however, at least in the relevant sense. The SAI might find them persuasive as analyses of the concept of "knowledge" or "justification" that is widely employed in everyday life, but there are clearly more rigorous concepts of "knowledge" and "justification" in existence (for instance, the ones employed by philosophers in their debates about deep philosophical issues). The SAI will surely not fail to grasp these concepts, and the SAI may judge them to be what matters in its own practical decision-making. But admittedly much remains up in the air here, since at this stage humans have so little understanding of what a feasible SAI would be like psychologically.

Objection (D) – Human philosophers have already solved all of these skeptical problems, or at least given convincing reasons to think that the skeptical problems are not worth worrying about. Consequently, premise (1) is false – an SAI will simply absorb the work of human philosophers and go about acting.

Response – Obviously it is impossible here to survey the history of epistemology and the various attempts to solve or diffuse skeptical problems. The field of problems and proposed solutions is probably too vast for a lifetime of philosophical work to be enough even for a relatively cursory treatment. If the problems have been solved, then that will be good news in our efforts to make progress in answering questions about SAI action. A handful of brief remarks is in order, though: first, proposals will need to be evaluated on a case by case basis. Given the diversity of skeptical problems and their suggested solutions, there is unlikely to be a general recipe that all the promising proposals will share. Insofar as this investigation has not already happened, it would be a useful contribution to answering the question of SAI action ultimately. Second, we must keep in mind once again that the SAI's path to action is potentially fragile – it might take only one skeptical hurdle that the SAI fails to clear for paralysis to result. Thus, depending on exactly what indirect approach the SAI is following, even solving large numbers of significant skeptical problems might not ensure a clear path to action. Finally, we must be wary of whatever is packed into locutions like "not worth worrying about" – i.e., we must be wary of proposals that are not solutions to these problems, but instead aim to give therapeutic responses. These therapeutic responses are often aimed at calming human worries in the face of skeptically driven angst, or explaining why humans do not wind up paralyzed and angst-ridden when it appears that indecision and angst are the rational responses. (Exactly how they are supposed to calm our worries, or explain why we don't have worries, without providing a solution to the problems varies from case to case.) Insofar as they succeed, they make us comfortable in our epistemological shoes because of (or because of their insights into) idiosyncratic aspects of human psychology – aspects that might not be shared by an SAI's psychology.

Objection (E) – If we are concerned that the SAI will fail to act because of skeptical problems, the designers of the SAI can alleviate the worry by ensuring that the SAI has determinate a priori degrees of belief in all salient propositions. These degrees of belief can be assigned in a way that respects our intuitive, common sense human views about the ceteris paribus plausibility orderings of all the various empirical hypotheses and philosophical principles. It would also be coherent and respect all the appropriate axioms of probability theory. When the SAI updates those degrees of belief based on evidence, the designers would also ensure that it follows commonly accepted principles of confirmation theory and decision theory – Bostrom flirts with such a proposal at one point (2014, 224–25). (For instance, the designers could ensure that the SAI's a priori degree of belief in the hypothesis that it is in a world of stable physical objects was much higher than its a priori degree of belief in the hypothesis that it is located in a simulation. Thus, when the evidence came in and was equally expected given either hypothesis, the SAI would ultimately favor the common sense physical object hypothesis.) This would make premise (6) false, because hard-wiring the SAI's initial priors in this way would give it essentially the same kinds of idiosyncrasies as human beings.

Response – Artificially giving the SAI such a system of a priori degrees of belief would be a massive undertaking. Even supposing that the undertaking could be carried out, there are further issues, some of which are by now likely to be familiar.[26] First, there are skeptical worries that might not themselves be resolved in the SAI just by ensuring that it had these determinate a priori degrees of belief, even assuming it never turned its philosophical attention on the degrees of belief. Such an SAI might continue to have skeptical worries in its attempts to investigate a variety of philosophical or mathematical propositions, for instance. The project of assigning a priori degrees of belief and then updating sounds like a paradigmatically Bayesian maneuver, and it may be impossible to apply Bayesian principles to philosophical or mathematical investigations. (I set aside here unusual and tricky exceptions that involve a posteriori investigations of paradigmatically a priori matters, such as when one polls mathematicians in an effort to learn if a particular theorem is true.)

Second, it is far from clear that such an agent would qualify as an SAI. It would essentially amount to an implicit, unreflective subjective Bayesian, because it would take its a priori credence assignment for granted and never turn any critical scrutiny on it.[27] (In fact, it would probably have to be unreflective about philosophical issues surrounding confirmation as a whole, since it would be very difficult given its massive cognitive ability for it to achieve reflectiveness about the issues generally but fail to draw obvious connections between its own practice and the philosophical debates.) It is plain that being an unreflective subjective Bayesian does not seem cognitively superior to considering the philosophical merits of subjective Bayesianism vis-à-vis competing positions. Yet plenty of human philosophers manage at least this much and arguably quite a bit more. Thus, such an "SAI" would have marked cognitive deficiencies even in comparison with many humans. It would have gaping blind spots in its cognition, much the way that current cutting edge AI programs, such as Watson or Deep Blue, have gaping blind spots in their cognition, impressive though they may be in certain narrow domains.[28] A very impressive artificial intelligence that has profound gaps in its theoretical reasoning that even moderately intelligent humans manage to avoid is not an SAI, because it is not systematically more cognitively skilled than human beings.

I acknowledge that whether an artificial intelligence with serious cognitive blindspots deserves the "SAI" moniker is ultimately just a linguistic matter, however. I prefer to define "superintelligence" as requiring systematic cognitive superiority to humans.[29] Some may prefer other definitions. It is, however, worth noting that individuals who countenance an understanding of SAIs that allows them to have serious cognitive deficiencies (equal to or surpassing those of many humans) will be forced to accept the possibility of "savant" SAIs – that is, SAIs that have much more in common with advanced versions of Deep Blue or Watson than with angels or gods. These agents might wind up ruling or destroying us, and they would undoubtedly dwarf us in certain mental respects, but they would not have surpassed us in all of our cognitive glory.

For the purposes of argument, I am happy to grant a less stringent understanding of the requirements for superintelligence. A final issue with the proposal here still emerges, though, and it is much more substantive. How feasible is it that something that would have the massive computational power, memory, and creativity to achieve the cognitive heights we envision from an SAI could also manage to be so bull-headed in its consideration of epistemological issues? This would require very impressive quarantining of the SAI's intellectual powers. This sort of quarantining may not be plausible given the architecture such an SAI would need to have to accomplish the wide variety of other feats that we generally take for granted that it would accomplish.

We must keep in mind that any SAI that vastly exceeded intelligent humans in a wide variety of domains (even if not in every domain) would almost undoubtedly have a very complex psychology. We must avoid the temptation to assume that it would function like a very fast but very unreflective instrumental rationality engine or Bayesian calculator. (Obviously, the greater its systematic advantage over humans, the more implausible these blindspots would become.) If we are to gain insights into whether quarantining is plausible (or even minimally likely), we must be very sensitive to the architectural details that a feasible SAI could have, and we must do our best to try to understand how those architectural details would affect the SAI's psychology. As of yet, there has not been much work on these issues. (The same basic points would apply for other similar strategies – e.g., artificially implanting the SAI with fallback beliefs to which it would automatically default in the event it gets stuck in a skeptical quandary.)

Objection (F) – Throughout the consideration of these objections (and the overall discussion of how skeptical worries might impact an SAI), you have repeatedly helped yourself to claims and arguments that make sense only if many or all of the skeptical worries you complain about are resolved. For instance, you make empirical claims about what is likely. These don't make any sense if some of the skeptical problems you discuss are real worries, as you seem to imply they are. Aren't you at best being hypocritical by talking this way, and aren't you at worst making it clear that you don't really believe that these skeptical problems are unsolved after all?

Response – I grant that I have used arguments that employ premises that are unjustified (in a rigorous sense of "justified"), assuming that we do not possess answers to the skeptical questions I have alluded to. I also grant that I do not have ready answers to at least many of these skeptical questions. My discussion in this section has been in the spirit of, loosely speaking, doing the best I can under the circumstances. I am just trying to come to insights in the way that humans normally do when they are trying to investigate empirical matters, particularly ones not easily amenable to controlled experiment or extensive evidence gathering.[30] It is true that the omnipresence of looming skeptical worries might force us in the end to remain agnostic about what an SAI would be like and how it would behave, just as these same skeptical worries might force us to withhold opinion about whether the world was created five minutes ago or we are brains in vats. And of course there are plenty of more humdrum considerations that justify intellectual humility in this specific instance – such as the widely appreciated problem of predicting technological developments and the special worry of understanding the ramifications of entities as complex as superintelligent AI. I offer up my arguments and speculations in the spirit of what humans normally do when they investigate the world, unjustified though their practices might ultimately be.

Conclusion

There are basically two principal issues our discussion has raised that are crucial for ultimately resolving questions about SAI action.

First, we must investigate how likely it is that SAIs will solve the host of skeptical problems that they may face in their philosophical and empirical investigations. While directly trying to solve the problems for ourselves in an effort to see how difficult they are may help, the history of philosophy gives us plenty of indirect grounds for pessimism. An approach that may bear fruit is trying to understand whether skeptical problems of various sorts are likely to be amenable to solutions that involve deployment of massive computational resources. Do these problems seem hard to us because we are not smart enough to solve them, or do they seem hard to us because they are unsolvable in principle?

Second, it will be important to examine how plausible it is for an SAI to have cognitive powers that are quarantined – the SAI manages to employ them to attain almost unimaginable heights in many intellectual domains while failing even to accomplish what humans do in others.[31] (In this case, it would not technically be an SAI, but nevertheless something that is cognitively superior in a wide variety of respects. As a result, it could thus pose existential risks.) Related to this is the issue of how plausible it is that the SAI would be more reasonable than the most ruthless of human investigators in some domains and as unreasonable as a toddler (or at least an unreflective human adult) in others. This will require us to understand in more detail than we do at present what form an SAI is likely to take, and how an SAI of that form's psychology will work. We will also need to be cautious about applying analogies from the human domain, since SAIs will probably be profoundly unlike humans in their psychology. (This is at least largely because they are likely to be engineered in a very different fashion, even if they have their origins in human neurobiology.)

As I discussed above, my aims in this paper have been mostly to raise issues for further work, not to conclusively settle any major debates. All of the issues I have raised admit of significant subtlety and complexity. I hope that future investigation is able to make further progress and apply the progress that's already been made to the case of superintelligent AI.

In the meantime, caution seems prudent. Although the motivational paralysis of an SAI is an underappreciated possibility, taking reckless risks with AI technology is not a good idea. Things could still go very badly, and the designer of a superintelligent AI need only succeed once to wreak havoc – AI could even destroy us before it ever reaches the superintelligence stage.

Notes

1. By "systematically" here, I mean possessing greater cognitive skill in every domain (or at least every important domain), not just in isolated areas. Also, whenever I use the term "intelligence," I use it as a synonym for "cognitive skill." The same goes for "smart." Incidentally, some theorists do believe that SAIs could have a biological basis: artificial extension and rewiring of the human brain, for instance. Hence "non-biological" should be understood here to mean "far from completely biological."

2. My rough introduction here is merely meant to motivate the issues I discuss later in the paper. It is not intended as a substitute for a thorough, rigorous treatment of topics surrounding the potential behavioral paths artificial intelligences might take and the prospects for controlling those paths.

3. For ease of exposition, I consider here a single SAI. If readers believe that the process would be too complex and daunting for one SAI to achieve on its own, imagine a more drawn-out saga that involved two or three generations of SAIs, perhaps culminating in the production of numerous SAIs that work together.

4. This definition is very rough, and would undoubtedly need to be sharpened considerably before it was perfectly adequate. Hopefully it is sufficient for now to capture the rough intuitive idea. Ironically, the difficulty of providing a definition of the Direct Approach is illustrative of the problem with the Direct Approach itself.

5. It might be objected that, if the SAI is genuinely much more cognitively skilled than human beings, it will recognize that we have failed to give it instructions that match what we really intended, and so will adjust its behavior to match our real aspirations. The trouble is that we are addressing ultimate pre-set goals for an AI – these would be the things that supposedly ultimately drive all of the AI's instrumental reasoning. It would not be possible for the AI to question or revise such a goal. There are complications lurking here that will be taken up later, however.

6. It should also be noted that the idealized circumstances involved in IP approaches must be specifiable in purely empirical terms, and cannot smuggle in qualifications that would definitionally tie the wants or beliefs of the group to the truth about morality. (Otherwise the IP approach would effectively collapse into an IM approach.) For example, the idealizations cannot include things like "what humans would believe about morality if they knew the truth about

morality" or "what humans would want if they were omniscient and desired nothing but to do right."

7. There is more that can be said to sharpen this definition, but for our purposes it should be precise enough.

8. My characterizations here are admittedly rough and imprecise. They are designed to give readers an intuitive grasp of the basic nature of the problems, not to give them a fully nuanced characterization of the state of the art in debates about the issues involved.

9. Two ways that philosophers have attempted to circumvent this skeptical problem are by proposing rigorous internalist coherentist theories of justification (Bonjour 1985 is a classic example) or by settling for some less rigorous – generally at least partially externalist – understanding of what would justify beliefs. Rigorous coherentism, however, suffers from a variety of skeptical problems of its own – described in Bonjour 1985. I will take up the issue of externalist theories of justification below. For the moment, suffice it to say that I am not convinced that an SAI would adopt such a view – at least not in a way that would help it to act in the face of the original skeptical worries.

10. Typically, this generalization is framed as a generalization from the past to the future, but it need not necessarily take that form. I will focus primarily on generalizations in time, however, since these are the most straightforward to analyze.

11. I place "region" in scare-quotes here because I am using the term loosely. It could be a literal spatio-temporal region, or it could be merely a bunch of cases in the general population of cases of interest.

12. Aside from ones guaranteed logically by the observed cases. For instance, if I have already observed Swan A at time *t* to be white, then any overall theory of the broader population that does not have Swan A at time *t* as white will be problematic.

13. I have in mind the version of the New Riddle of Induction in Goodman 1955, although it is expressed differently there from how it is in more recent debates that use Bayesianism as an organizing framework. Fortunately, the details are not relevant for our purposes. For those used to approaching issues of evidential reasoning from a Bayesian perspective, one can appreciate the skeptical problem by asking how the initial priors (i.e., the a priori, intrinsic priors) for the various hypotheses should be set.

14. I do not refer to these epistemic probabilities simply as "prior probabilities" because they are prior to all evidence, not just the current evidence. The terminology of a "prior probability" can sometimes be ambiguous between the two options.

15. An SAI following an IM might luck out if the version of moral realism that winds up being true directly entails specific actions that should be done, rather than spelling out various empirical features that actions must have in order to be right. Very few people believe in such a form of moral realism, however, for whatever that is worth. Arguably, even pure Kantianism would not qualify. And there might still be varieties of a priori skepticism that must be faced**.**

16. In the interest of simplicity, I assume throughout that any true moral theory would be anti-supererogationist. The issues that could potentially plague the SAI don't change fundamentally even if supererogationism is true, but there is need for added complexity in specifying how the

SAI would choose between permissible actions. For some related discussion, see Bostrom 2014, chapter 13.

17. Although I will not address them here, there might even be more global and insidious varieties of skepticism that the SAI struggles with and ultimately fails to resolve, such as the Kripkenstein problem. See the famous discussion in Kripke 1980, especially chapter 2.

18. See below for a discussion of the deliberate engineering of SAIs to have cognitive deficiencies. Incidentally, the deliberate engineering need not be direct. It could involve some kind of evolutionary process that imposed selection pressures which rewarded these deficiencies.

19. More must be said undoubtedly to make the idea of a cognitive failure perfectly precise, but nothing I say here will trade on any objectionable details of formulation.

20. It may be that use can be made here and elsewhere of the Alief/Blief distinction (see, for instance, Szabó Gendler 2008). There are also matters that must be addressed regarding the relationship between conscious beliefs and standing subconscious beliefs. I will not speculate further, as any such discussion will lead to subtle issues that are well beyond the scope of what I can hope to accomplish in this paper. Also, if it is technically wrong to speak of full belief here, simply substitute "high level of confidence" or "high degree of belief" where relevant.

21. I am assuming a picture of action where action requires some fundamental desire or moral belief (where "desire" has a minimalist interpretation that does not smuggle in inappropriately anthropomorphic assumptions). One then forms empirical beliefs connecting this fundamental moral belief or desire with action. (In the process, this may cause other more specific moral beliefs and instrumental desires to form. The precise details depend on subtle philosophical and empirical issues that we need not worry about here.) So, for instance, I might have a fundamental desire for pleasure, believe that riding my bike will give me pleasure, and believe that grabbing my bike and sitting on it will be the best way to start the process of riding it. I then execute this action.

22. I ignore here the issue of whether an IP SAI would think about philosophical issues at the foundations of morality, and I presuppose that any failures to answer such questions would not somehow undermine the pre-programmed ultimate desire of the SAI. These questions merit further attention, however. Also, none of what I have said here about the difficulties for IM SAIs should be taken to imply that there would be smooth sailing for IP SAIs. Their attempts to discern the concrete content of their final goals would still depend at least on their attempts to deal with empirical skeptical worries. More on this below.

23. It is unclear exactly what role reasonableness plays here (i.e., attentiveness and responsiveness to rational considerations about justification) and what role cognitive skill plays directly in allowing smarter individuals to see potential problems and subtleties that less smart individuals miss.

24. For those accustomed to thinking in a Bayesian fashion, the issue here has to do with setting a principled value of P(h) for each hypothesis, prior to receiving any evidence.

25. There are some technical qualifications that may apply to theories like Sepielli's, but none of those qualifications will impact the fundamental points I am making. I also set aside issues about countable additivity.

26. One way the designers might successfully carry out the undertaking is indirect – via exposing predecessor AIs to an evolutionary process with selection pressures that rewarded action. The speculations below can then be seen as calling into question the idea that one could produce an SAI using this sort of evolutionary process.

27. For our purposes, subjective Bayesianism is roughly the view that agents should update prior degrees of belief in light of the evidence using Bayesian principles, plus the claim that any assignment of a priori degrees of belief that is consistent with the axioms of probability theory is justified. In other words, no such assignment of a priori degrees of belief is superior from a justificatory standpoint to any other.

28. Watson and Deep Blue are IBM computers designed to play Jeopardy! and chess respectively. Deep Blue famously defeated world champion Gary Kasparov in a chess series in 1997. Watson won a multi-game Jeopardy! competition in 2011 against Brad Rutter and Ken Jennings, the two biggest winners in the *Jeopardy!* television show's history.

29. This is quite close to Bostrom's definition: a superintelligence is "any intellect that exceeds the cognitive performance of humans in virtually all domains of interest" (2014, 22 – one gets the impression that the only domains where such an agent would not exceed the cognitive performance of humans are ones where humans are already nearly topping out the scales). It is also basically equivalent to I.J. Good's classic definition of "ultraintelligence": an ultraintelligent machine is "a machine that can far surpass all the intellectual activities of any man however clever" (Good 1965, 33).

30. Although I have addressed many distinctively philosophical issues in this section – issues relating to skepticism and rationality – the matter here is ultimately empirical, because the question is how a particular kind of artificial cognitive agent, constructed of tangible building blocks, will address these philosophical issues and behave in response.

31. Obviously some computers already do this in a few narrow domains. But we are speaking here of artificial intelligence that exceeds human capacity in a wide variety of domains, including domains where computers have thus far demonstrated little aptitude.

Acknowledgments

Thanks to Audre Brokes, Susan Schneider, Gia Montemuro, and audiences at Leverage Research and the University of Lisbon for their helpful feedback on previous versions of this paper.

References

Bonjour, L. 1985. *The structure of empirical knowledge*. Cambridge, MA: Harvard University Press.

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Chalmers, D. 2012. The singularity. *Journal of Consciousness Studies* 17: 7–65.

Devries, W., and M. Triplett. 2000. *Knowledge, mind, and the given*. Indianapolis: Hackett.

Goldman, A. 1976. What is justified belief?. In *Justification and knowledge,* ed. G.S. Pappas, 1–25. Dordrecht: Reidel.

Good, I.J. 1965. Speculations concerning the first ultraintelligent machine. In *Advances in computers*, ed. F.L Ault and M. Rubinoff, 31–88. New York: Academic Press.

Goodman, N. 1955. *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Hume, D. 1978. *A treatise of human nature*. Ed. L.A. Selby-Bigge and P.H. Nidditch. 2nd ed. Oxford: Clarendon Press. (Orig. pub. 1738.)

Kripke, S. 1980. *Wittgenstein on rules and private language*. Cambridge, MA: Harvard University Press.

Mackie, J.L. 1977. *Ethics: Inventing right and wrong*. London: Penguin.

Resnik, M. 1987. *Choices: An introduction to decision theory*. Minneapolis: University of Minnesota Press.

Russell, B. 1918. The philosophy of logical atomism. *The Monist* 28: 495–527.

Sellars, W. 1956. Empiricism and the philosophy of mind. In *Minnesota Studies in the Philosophy of Science*, vol. I, ed. H. Feigl and M. Scriven, 253–329. Minneapolis: University of Minnesota Press.

Sepielli, A. 2009. What to do when you don’t know what to do. *Oxford Studies in Metaethics* 4: 5–28.

Sepielli, A. 2013. What to do when you don’t know what to do when you don’t know what to do… *Nous* 47 (1): 521–44.

Szabó Gendler, T. 2008. Alief and blief. *Journal of Philosophy* 105(10): 634–63.