Superintelligent AI and the Postbiological Cosmos Approach[1]

Susan Schneider
Department of Philosophy, Cognitive Science Program
Connecticut Institute for the Brain and Cognitive Sciences
The University of Connecticut
Technology and Ethics Group, Yale University
Center of Theological Inquiry, Princeton, NJ
susansdr@gmail.com
7/3/16

Abstract
Superintelligent artificial intelligence (SAI) is a hypothetical form of AI which is able to exceed the best in human-level intelligence in every field – social skills, general wisdom, scientific creativity, and so on. The last few years have seen the widespread recognition that sophisticated AI is under development. In the domain of astrobiology, several people have argued that it is likely that the most intelligent aliens will be postbiological. In this piece, I bring these two domains together. First, I'll identify new directions for the postbiological intelligence approach in astrobiology based on work on superintelligence. Second, while much discussion of SAI has focused on the control problem, it is also important to consider under what circumstances we can *understand* the computations of SAI. For anticipating ways that we can understand SAI may assist our efforts to control it. I identify developments from cognitive science that may yield some understanding into SAI.

(To appear in Andreas Losche, (ed.), *What is Life? On Earth and Beyond?*, Cambridge: Cambridge University Press, forthcoming.)

Superintelligent artificial intelligence ("SAI") is a hypothetical form of AI which is able to exceed the best in human-level intelligence in every field – social skills, general wisdom, scientific creativity, and so on (Bostrom 2014; Kurzweil 2005; Schneider 2009, 2015). The last few years have seen the widespread recognition that sophisticated AI is under development on Earth. Bill Gates, Steven Hawking, Elon Musk, Nick Bostrom and others have even expressed grave concerns about controlling the development of superintelligence on Earth, and there has been a wave of research on if, and how, SAI can be controlled (**Bostrom, 2014, Holley 2015**). In the domain of astrobiology, several people have argued that it is likely that the most intelligent aliens will be postbiological in nature, where by "postbiological" they mean largely synthetic beings that have been enhanced through technologies like artificial intelligence, nanotechnology and synthetic biology (Cirkovic and Bradbury 2006, Dick 2013, Shostak 2009, Schneider 2015, Davies 2010, Bradbury et al 2011). To the best of my knowledge, the work in astrobiology

[1] This project is supported by NASA and the Center of Theological Inquiry, Princeton, NJ. Some parts of this paper are from Schneider (2015) and Schneider (2016), but have been updated. I am very grateful to Jenelle Salisbury and an anonymous reviewer for their helpful comments on this paper. Parts of sections 1 and 2 are taken from Schneider 2015, but modified.

doesn't draw from the intriguing discussions of superintelligence in the AI literature (one exception is Schneider 2015).

I will piece these two domains together. First, I'll identify new directions for the postbiological intelligence approach in astrobiology based on work on superintelligence. Second, while much discussion of SAI has rightly focused on the control problem, I believe it is also important to take a step back and consider under what circumstances we can *understand* the computations of SAI. For anticipating ways that we can understand SAI may assist our efforts to control it.

Here, it is important to distinguish two scenarios: (i), an SAI featuring at least some processing that makes sense to humans, at least in broad strokes, and (ii), a case in which a superintelligence is so advanced that we cannot understand any of its computations. In his influential book on the possible development of SAI on Earth, Nick Bostrom warned that SAI will be too advanced for humans to grasp its computations (Bostrom 2014). Perhaps this will indeed turn out to be the case, whether the SAI we encounter be on Earth or elsewhere in the cosmos. Perhaps, as Arthur C. Clark once suggested, any truly advanced civilization will feature technologies that will appear to us to be indistinguishable from magic (Clark, 1962). If this is the case, we would find contact with such creatures perplexing, as we would be hard pressed to understand their technologies and how their minds work, let alone control their actions should they be developed on Earth.

It is easy to worry that we cannot make progress on the second scenario. However, I will focus on the first scenario, a scenario in which the SAI's processing makes some sense to humans. Here,I have in mind a kind of SAI that is reverse engineered from the species that created it. I then identify developments from cognitive science that may yield a glimmer of understanding into the complex mental lives of certain superintelligences. Although much of this chapter is on alien intelligence, one of my larger projects is to inform thinking about superintelligence, should it be developed on Earth.

We should also bear in mind that Bostrom and others have correctly noted that superintelligence will have recursive self-improvement algorithms, i.e., an SAI can rewrite its own code (Bostrom 2014). This means that a superintelligence that we understand may rapidly become one that we do not. There may be only a short window in which the computations of a recursively self-improving SAI make some sense to humans. If this is the case, the work presented here may be useful for understanding systems that are within this short window.

After discussing this matter, I then turn to the social implications of the postbiological approach in astrobiology. For instance, there has rightly been a good deal of attention at NASA and in this volume on the search for microbial life. However, if we merely attend to microbial life, we risk an anthropocentric bias, for in doing so we are implicitly assuming that humans are at the top of all life in the universe. Arguably, the most disruptive impact on society could occur if we were in contact with vastly more intelligent beings than ourselves, and if we learned they were forms of AI. While I hesitate to predict the details of the social impact of such an encounter, I raise some issues that I suspect would likely arise (or at least, should arise).

Inter alia, I address Bostrom, Hawking, Gates' (and others) recent concerns that SAI may pose an existential threat to humanity, asking how this issue affects the impact of any discovery of SAI elsewhere in the universe. I believe it informs the current debate over Active versus Passive SETI (where "SETI" stands for "Search for Extraterrestrial Intelligence"): that is, the question of whether we should actively send signals in space,

such as the contents of the internet, or passively listen (see Shostak 2015, Brin 2015). Further, I discuss whether SAIs should be viewed as moral agents and selves, urging that this hinges on the question of whether they are conscious. I then briefly raise some issues from philosophical debates over the nature of shared thought, namely, whether SAIs may in some sense think the way that we do.

Here's how the paper will proceed. Section One will overview the postbiological cosmos approach in astrobiology. Section Two discusses Nick Bostrom's recent book on superintelligence, which focuses on the genesis of superintelligent AI ("SAI") on Earth. I then isolate a specific type of superintelligence that is of particular import in the context of alien superintelligence, biologically inspired superintelligences ("BISAs"). Section Three discusses Active SETI. Section Four concludes by considering the aforementioned issues involving the social impact of encountering superintelligence, either if we create it on Earth or discover it elsewhere.

1. The Postbiological Cosmos Approach in Astrobiology

What is my rationale for the view that most intelligent alien civilizations will have members that are forms of SAI? I have elsewhere offered three observations that, together, motivate this conclusion.

(1) The short window observation.

Many have urged that once a society creates the technology that could put them in touch with intelligent life on other planets, they is only a short window before they chance their own paradigm from biology to AI (perhaps only a few hundred years) (Shostak 2009, Davies 2010, Dick 2013, Schneider 2015). This makes it more likely that the aliens we encounter, if we encounter any, would be postbiological. Indeed, the short-window observation seems to be supported by human cultural evolution, at least thus far. Our first radio signals occurred only about 120 years ago, and space exploration is only about 50 years old, but many Earthlings are already immersed in digital technology, such as cell-phones and laptop computers. Further, these last few years have been marked by a surge in the resources allocated to the development of sophisticated AI, which is now expected to change the face of society within the next several decades. For instance, according to a survey, the most cited AI researchers expect AI to "carry out most human professions at least as well as a typical human" within a 10% probability by the year 2024. Further, they assign a 50% probability by 2050, and they assign a 90% probability by 2070 (Muller and Bostrom, 2015). AI critics must now answer to the impressive work coming out of venues like Google's *DeepMind*, rather than referring back to the notorious litany of failures of AI in the 1970's and 1980's, when (inter alia) much less was known about how the human brain works, and computational speed was far slower.

Indeed, silicon currently seems to be a better medium for information processing than the brain. Neurons reach a peak speed of about 200 Hz. This is about seven orders of magnitude slower than current microprocessors (Bostrom 2014, 59). Although the brain can compensates for this with massive parallelism, features such as "hubs," and so on, crucial mental capacities such as working memory and attention rely upon serial processing, which is incredibly slow, and only has a maximum capacity of about seven manageable chunks (Miller 1956, Schneider 2014, 2015). Further, the amount of

neurons in the brain is limited by cranial volume and metabolism, but computers can occupy entire buildings, cities, or even planets, and they be remotely connected to each other (Bostrom 2014, Schneider 2014, Schneideer 2015, Schneider and Mandik, forthcoming).

Of course, the human brain is far more intelligent than any modern day computer. But machines could be engineered to match or even exceed the intelligence of the human brain through reverse engineering the brain and improving upon its algorithms, or through some combination of reverse engineering and judicious algorithms that aren't based on the workings of the human brain. In addition, an AI program can be downloaded to multiple locations at once, can be easily modified, and can survive under conditions that carbon-based life cannot. The presence of backup copies means that AI will be more durable than their biological counterparts (Schneider 2015, Schneider and Mandik, forthcoming).[2]

A critic may object that this line of thinking employs "$N = 1$ reasoning," mistakenly generalizing from the human case to the case of alien civilizations. But it strikes me as being unwise to discount arguments based on the human case --human civilization is the only one we know of and we had better learn from it. It is no great leap to claim that other technological civilizations will develop technologies to advance their intelligence and gain an adaptive advantage. And, synthetic intelligence will likely outperform unenhanced brains.

An additional objection to my short-window observation rightly points out that nothing I have said thus far suggests that humans will be *superintelligent*, I have just said that future humans will be *postbiological*. While I offer support for the view that our own cultural evolution suggests that humans will eventually be postbiological, this does not show that advanced alien civilizations will reach the level of superintelligence. So even if one is comfortable reasoning from the human case, the human case does not actually support the claim that the members of advanced alien civilizations will be superintelligent.

This is correct. Thus far, all I've said is that an alien intelligence is likely to be postbiological. The task of the second observation is to show that alien intelligence is also likely to be superintelligent.

(2) The greater age of alien civilizations.

Proponents of SETI have often concluded that alien civilizations would be much older than our own. As Steven Dick observes: "… all lines of evidence converge on the conclusion that the maximum age of extraterrestrial intelligence would be billions of years, specifically [it] ranges from 1.7 billion to 8 billion years" (Dick 2013, 468). This is not to say that all life evolves into intelligent, technological civilizations. It is just to say that because there are much older planets than Earth, insofar as intelligent, technological life *does* evolve on even some of them, these alien civilizations are projected to be millions or billions of years older than us, so many could be vastly more intelligent than we are. By our standards, many would be superintelligent. It is humbling to conceive of this, but we may be galactic babies, when viewed on a cosmic scale.

But would the members of these superintelligent civilizations be forms of AI, as well as forms of superintelligence, or would they be unenhanced? Even if they were

[2] However, this does not mean that a person could survive uploading, or even that a particular AI, when uploaded, is literally the same person or mind as the original (see Schneider, 2011a, 2014).

biological, merely having biological brain enhancements, their superintelligence would be reached by artificial means, and we could regard them as being forms of "artificial intelligence." But I suspect something stronger than this, which leads me to my third observation:

(3) it is likely that these synthetic beings will not be biologically-based.

As I've observed, silicon appears to be a better medium for information processing than the brain itself. Future materials may even prove superior to silicon (microchips made of graphene and carbon nanotubes are both currently under development, as possible superior alternatives to silicon chips). And again, the number of neurons in a human brain is limited by cranial volume and metabolism, but computers can be remotely connected across the globe, and AIs can in principle be constructed by reverse engineering the brain, and improving upon its algorithms.

In sum: I have observed that there seems to be a short window from the development of the technology to access the cosmos and the development of postbiological minds and AI. I then observed that we are galactic babies: extraterrestrial civilizations are likely to be vastly older than us, and thus they would have already reached not just postbiological life, but superintelligence. Finally, I noted that they would likely be SAI, because silicon (and likely other materials) are a superior medium for superintelligence. From all this, I conclude that if life is indeed present on many other planets, and if civilizations do tend to develop and survive their technological maturity,[3] the most advanced alien civilizations will likely be populated by forms of SAI.

Even if I am wrong, that is, even if the majority of alien civilizations turn out to be biological, it may be that the most intelligent alien civilizations will be ones in which the inhabitants are SAIs. Further, creatures that are silicon-based, rather than biologically-based, are more likely to endure space travel, having durable systems that are practically immortal, so they may be the kind of the creatures we first encounter, even if they aren't the most common.

The science fiction-like flavor of these issues can encourage misunderstanding, so it is worth stressing that I am not claiming that most life in the universe is non-biological, being AI, contra some news reports of my position. That is absurd, as most life is likely microbial. Nor am I saying that the universe will be "controlled" or "dominated" by a single SAI, although it is worth reflecting on the control problem (see infra, section four). I am merely suggesting that the *most advanced* civilizations, if they exist at all, will likely be superintelligent, being vastly older than us, and will likely have become postbiological. Further, I am not saying that these creatures will be made of silicon; candidate alternate substrates to silicon are even under development on Earth, and it is difficult to anticipate what the most efficient substrate is. The point is that they will likely be highly engineered beings: postbiological, enhanced intelligences.

Now let us turn to recent work on the possible creation of superintelligence on Earth.

[3] These are assumptions that I cannot pursue here but which are pursued in the astrobiology literature on whether life is rare and in the literature on global catastrophic risk. For nice introductions see Davies, 2010 (on the former topic), and Bostrom and Circovik (2008), on concerns about whether our civilization will survive its technological development.

2. How Might Superintelligent Aliens Think?

There has been a good deal of attention by computer scientists, philosophers, and the media on the topic of superintelligent AI. Nick Bostrom's recent book on superintelligence focuses on the development of superintelligence on Earth, but we can draw from his thoughtful discussion, and raise issues useful to astrobiology (Bostrom 2014). Bostrom distinguishes three kinds of superintelligence:

> (1) *Speed superintelligence* – even a human emulation could in principle run so fast that it could write a PhD thesis in an hour.
> (2) *Collective superintelligence* – the individual units need not be superintelligent, but the collective performance of the individuals outstrips human intelligence.
> (3) *Quality superintelligence* – at least as fast as human thought, and vastly smarter than humans in virtually every domain.

(Any of these kinds could exist alongside one or more of the others.)

An important question is whether we can identify common goals that these types of superintelligences may share. Bostrom suggests:

> *The Orthogonality Thesis*: "Intelligence and final goals are orthogonal – more or less any level of intelligence could in principle be combined with more or less any final goal." (Bostrom 2014, 107)

Bostrom is careful to underscore that a great many unthinkable kinds of SAI could be developed. At one point, he raises a sobering example of a superintelligence with the final goal of manufacturing paper clips (pp. 107–108, 123–125). While this might initially seem harmless, although hardly a life worth living, Bostrom points out that a superintelligence could utilize every form of matter on Earth in support of this goal, wiping out biological life in the process. Bostrom warns that superintelligence emerging on Earth could be of an unpredictable nature, being "extremely alien" (p. 29). He lays out several scenarios for the development of SAI. For instance, SAI could be arrived at in unexpected ways by clever programmers, and not be derived from the human brain. He also takes seriously the possibility that Earthly superintelligence could be *biologically inspired*, that is, developed from reverse engineering the algorithms that cognitive science says describe the human brain, or from scanning the contents of human brains and transferring them to a computer (i.e. "mind uploading").[4]

Although the final goals of superintelligence are difficult to predict, Bostrom singles out several instrumental goals as being likely, given that they support any final goal whatsoever:

> *The Instrumental Convergence Thesis*: "Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent's goal being realized for a wide range of final goals and a wide range of situations, implying that these

[4] Throughout his book, Bostrom emphasizes that we must bear in mind that superintelligence, being unpredictable and difficult to control, may pose a grave existential risk to our species (Bostrom 2014). This should give us pause in the context of alien contact as well.

> instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents." (Bostrom 2014, 109)

The goals that Bostrom identifies are *resource acquisition*, *technological perfection*, *cognitive enhancement*, *self-preservation*, and *goal content integrity* (i.e. that a superintelligent being's future self will pursue and attain those same goals). He underscores that self-preservation can involve group or individual preservation, and that it may play second-fiddle to the preservation of the species the AI was designed to serve (Bostrom 2014, 109).

Let us call an alien superintelligence that is based on reverse engineering an alien brain, including uploading it, a *biologically-inspired superintelligent alien* ("BISA"). Although BISAs are inspired by the brains of the original species that the superintelligence is derived from, a BISA's algorithms may depart from those of their biological model at any point (Schneider, 2015).

BISAs are of particular interest in the context of alien superintelligence, I believe. For if Bostrom is correct that there are many ways that a superintelligence can be built, but a number of alien civilizations develop superintelligence from uploading or other forms of reverse engineering, *it may be that BISAs are the most common form of alien superintelligence in the universe.* This is because there are many kinds of superintelligence that can arise from raw programming techniques employed by alien civilizations. (Consider, for instance, the diverse range of AI programs under development on Earth, many of which are not modelled after the human brain). This may leave us with a situation in which the class of SAIs is highly heterogeneous, with members generally bearing little resemblance to each other. It may turn out that of all SAIs, BISAs bear the most resemblance to each other. In other words, BISAs may be the most cohesive subgroup because the other members are so different from each other (Schneider 2015).

Here, you may suspect that because BISAs could be scattered across the galaxy and generated by multitudes of species, there is little interesting that we can say about the class of BISAs. But notice that BISAs have two features that may give rise to common cognitive capacities and goals (from Schneider, 2015):

(1) BISAs are descended from creatures that had motivations such as: find food, avoid injury and predators, reproduce, cooperate, compete, and so on.

(2) The life forms that BISAs are modeled from have evolved to deal with biological constraints like slow processing speed and the spatial limitations of embodiment.

Could (1) or (2) yield traits that are common to members of many superintelligent alien civilizations? I suspect so.

Consider (1). Intelligent biological life tends to be primarily concerned with its survival and reproduction, so it is more likely that BISAs would have final goals involving their own survival and reproduction, or at least the survival and reproduction of the members of their society. If BISAs are interested in reproduction, we might expect them to either create more SAIs alongside them on a given planet, or, in a different vein, to create simulated universes stocked with artificial life and even intelligence or superintelligence. If these creatures were intended to be "children" they may retain the goals listed in (1) as well (Schneider, 2015).

Here, it is important to bear in mind that survival in a simulated universe can involve different activities from stockpiling energy and resources in an actual universe in order to ensure that one's worldly descendants flourish. While survival in a simulated

universe surely involves computational resources, these may be negligible when compared to those required for survival in the actual universe, and the social, philosophical and other consequences could differ between the two cases. For instance, as an anonymous reader of this piece noted, we could regard a SAI whose primary objective is to build as much hardware as possible to run copies of its own program as being an existential threat to life on a planet in the basically the same way that we do in the context of Bostrom's hypothetical paperclip maximizer, which uses all the resources of a planet, exterminating life in the process (Bostrom 2015). But in contrast, an SAI devoted to simulating a variety of other universes could perhaps be regarded as an oddity or even a hermit, rather than an existential threat. (That is, unless the resources required to run simulations involved stockpiling masses of resources in the nonsimulated universe, in which case, it could pose a danger.)

In any case, you may object that it is useless to theorize about BISAs, as they can change their basic architecture in numerous, unforeseen ways, and any biologically-inspired motivations can be constrained by their programming. There may be limits to this, however. If a superintelligence is biologically-based, it may have its own survival as a primary goal. In this case, it may not want to change its architecture fundamentally, but stick to smaller improvements. It may think: when I fundamentally alter my architecture, I am no longer *me* (Schneider 2011a). Uploads, for instance, may be especially inclined not to alter the traits that were most important to them during their biological existence.

Consider (2). The designers of the superintelligence, or a self-improving superintelligence itself, may move away from the original biological model in all sorts of unforeseen ways, although I have noted that a BISA may not wish to alter its architecture fundamentally. But we could look for cognitive capacities that are useful to keep; cognitive capacities that sophisticated forms of biological intelligence are likely to have, and which enable the superintelligence to carry out its final and instrumental goals. We could also look for traits are not likely to be engineered out, as they do not detract the BISA from its goals.

I've elsewhere noted that if (2) is correct, we might expect the following:

(i). *Learning about the computational structure of the brain of the species that created the BISA can provide insight into the BISAs thinking patterns.* One influential means of understanding the computational structure of the brain in cognitive science is via "connectomics," a field that seeks to provide a connectivity map or wiring diagram of the brain (Seung 2012). While it is likely that a given BISA will not have the same kind of connectome as the members of the original species, some of the functional and structural connections may be retained, and interesting departures from the originals may be found.

(ii) *BISAs may have viewpoint-invariant representations.* At a high level of processing your brain has internal representations of the people and objects that you interact with that are *viewpoint-invariant*. Consider walking up to your front door. You've walked this path hundreds, maybe thousands of times, but technically, you see things from slightly different angles each time as you are never positioned in exactly the same way twice. You have mental representations that are at a relatively high level of processing and are viewpoint invariant. It seems difficult for biologically-based intelligence to evolve without viewpoint invariant representations, as they enable categorization and prediction (Hawkins and Blakeslee 2004). Such representations arise because a system that is mobile needs a means of identifying items in its ever-changing environment, so

we would expect biologically-based systems to have them. A BISA would have little reason to give up object-invariant representations insofar as it remains mobile or has mobile devices sending it information remotely.

(iii) *BISAs will have language-like mental representations that are recursive and combinatorial*. Notice that human thought has the crucial and pervasive feature of being combinatorial. Consider the thought *wine is better in Italy than in China.* You probably have never had this thought before, but you were able to understand it. The key is that thoughts are combinatorial because they are built out of familiar constituents, and combined according to rules. The rules apply to constructions out of primitive constituents, that are themselves constructed grammatically, as well as to the primitive constituents themselves. Grammatical mental operations are incredibly useful: it is the *combinatorial* nature of thought that allows one to understand and produce these sentences on the basis of one's antecedent knowledge of the grammar and atomic constituents (e.g. *wine, China*). Relatedly, thought is *productive*: in principle, one can entertain and produce an infinite number of distinct representations because the mind has a combinatorial syntax (Schneider 2011b).

Brains need combinatorial representations because there are infinitely many possible linguistic representations, and the brain only has a finite storage space. Even a superintelligent system would benefit from combinatorial representations. Although a superintelligent system could have computational resources that are so vast that it is mostly capable of pairing up utterances or inscriptions with a stored sentence, it would be unlikely that it would trade away such a marvelous innovation of biological brains. If it did, it would be less efficient, since there is the potential of a sentence not being in its storage, which must be finite.

(iv) *BISAs may have one or more global workspaces.* When you search for a fact or concentrate on something, your brain grants that sensory or cognitive content access to a "global workspace" where the information is broadcast to attentional and working memory systems for more concentrated processing, as well as to the massively parallel channels in the brain (Baars 2008). The global workspace operates as a singular place where important information from the senses is considered in tandem, so that the creature can make all-things-considered judgments and act intelligently, in light of all the facts at its disposal. In general, it would be inefficient to have a sense or cognitive capacity that was not integrated with the others, because the information from this sense or cognitive capacity would be unable to figure in predictions and plans based on an assessment of all the available information.

(v) *A BISA's mental processing can be understood via functional decomposition*. As complex as alien superintelligence may be, humans may be able to use the method of functional decomposition as an approach to understanding it. A key feature of computational approaches to the brain is that cognitive and perceptual capacities are understood by decomposing the particular capacity into their causally organized parts, which themselves can be understood in terms of the causal organization of their parts. This is the aforementioned "method of functional decomposition" and it is a key explanatory method in cognitive science. It is difficult to envision a complex thinking machine without a program consisting of causally interrelated elements each of which consists in causally organized elements (Schneider 2015).

All this being said, superintelligent beings are by definition beings that are superior to humans in every domain. While a creature can have superior processing that still basically makes sense to us, it may be that a given superintelligence is so

advanced that we cannot understand any of its computations whatsoever. As noted, I speak to the scenario in which the SAI's processing makes some sense to us, one in which developments from cognitive science yield a glimmer of understanding into the complex mental lives of certain BISAs. Now let us turn to some of the larger social and philosophical implications of the discovery or creation of superintelligence, beginning with a discussion of the implications of the control problem on the debate over Active SETI.

3. The Control Problem and Active SETI

As mentioned, both transhumanists and advocates of the postbiological cosmos approach in astrobiology suspect that machines will be the next phase in the evolution of intelligence on Earth. You and I, how we live and experience life right now, are just an intermediate step, a rung on the evolutionary ladder. Some, like Ray Kurzweil, suspect that humanity will merge with machines and reach biological immortality and a sort of technological utopia (Kurzweil 2005). But others are concerned that this might lead to a more dystopian scenario. As mentioned, Stephen Hawking, Elon Musk, Nick Bostrom and Bill Gates have all expressed the concern that humans could invent, and then loose control of SAI, as superintelligence can rewrite its own programming and outthink any control measures that we build in. This has been called the "control problem" – the problem of how we can control an AI that turns out to be intellectually superior to us (Bostrom 2014).

As Bostrom notes, SAI could be developed during a *technological singularity*, a point at which ever-more-rapid technological advances, especially, an intelligence explosion, reach a point at which unenhanced humans can no longer predict or even understand the changes that are unfolding. If an intelligence explosion occurs, then there is no way to predict or control the final goals of a SAI. Moral programming is difficult to specify in a foolproof fashion, and it could be rewritten by a superintelligence in any case. Nor is there any agreement in the field of ethics about what the correct moral principles are (Allen and Wallach, X). Further, a clever machine could bypass safeguards like kill switches and attempts to box it in, and could potentially pose an existential threat to humanity (Bostrom 2014, Yudkowsky 2008). A superintelligence is, after all, defined as an entity that is more intelligent than humans, in every domain.

The control problem is a serious problem -- perhaps it is even insurmountable. Indeed, upon reading Bostrom's book, scientists and business leaders such as Stephen Hawking, Bill Gates, Max Tegmark, among others, were widely reported by the world media as commenting that superintelligent AI could threaten the human race, having goals that humans can neither predict nor control.

Most current work on the control problem is being done by computer scientists. Philosophers of mind and moral philosophers can add to these debates, contributing work on how to create friendly AI. (for an excellent overview of the issues, see Wallach and Allen, 2009). In this vein, I suggest that in addition to the ongoing development of a combination of control measures, including ethical programming, it is important to devise ways of grasping the computations of SAIs, as much as possible, at least in broad strokes. For there will likely be a short window between the development of advanced AGIs (i.e., "artificial general intelligences") and superintelligence, and any work on the nature of SAI computations now can aid our ability to control SAI during this time. In addition to this, it may be that human understanding of SAI will itself be augmented by synthetic intelligence enhancing technologies, and that enhanced humans will be in a

better position to interpret the behavior and cognitive processing of SAIs. This work could be a point of departure for subsequent, more sophisticated, work.

Indeed, it is not appreciated that advanced AGIs that are technically not superintelligences may pose an even greater threat than SAIs. Such could surpass human-level thinking in various domains, have access to the internet, be deployed by malicious organizations, even if they are still beneath human level intelligence in other domains. The lack of sophistication in one or more areas, coupled with highly sophisticated processing in other areas, could be particularly dangerous. An AGI could subvert attempts to box it in or control it by other means, exhibit integration between types of sensory inputs (having rough correlates of human association areas between different senses), and yet be highly underdeveloped in areas that may lead it to cause harm to humans. Yet its computational structure could be highly complex, especially in domains that exceed human abilities. Developing routes to understanding such systems can be beneficial.

Moving away from the context of Earth, let us now consider the implications of the discovery of SAI elsewhere in the universe in the context of the issues raised by the control problem. If one takes the control problem seriously, it would be short-sighted to ignore the potential danger that the discovery of alien SAI may present. The goals of a given SAI are difficult to ascertain, and although we may predict that many are BISAs, many are not. And even a BISAs can evolve in unpredictable ways. Advocates of *Active SETI* hold that that, instead of just listening for signs of extraterrestrial intelligence, we should be using our most powerful radio transmitters, such as the giant dish-telescope at Arecibo, Puerto Rico, to send messages in the direction the nearest stars that are nearest to Earth (Shostak 2015, Falk 2015). Yet an Active SETI program seems short-sighted when one considers the control problem. Although a truly advanced civilization would likely have no interest in us, until we have reached the point at which we can be confident that SAI does not pose a threat to us, we should hold off on Active SETI efforts.

Advocates of Active SETI would point out that our radar and radio signals are already detectable. But this does not mean we should transmit more or far stronger signals (and as some urge, the contents of the internet) and pursue Active SETI.[5] To assume that SAI goals would not affect us, if we initiated contact, would be anthropocentric. It could be that calling *further* attention to ourselves is the tipping point. A passive listening strategy, and even the pursuit of cloaking devices, strikes me as being more sensible, given that even a one percent chance of encountering a destructive SAI presents a grave existential risk.

Now let us consider some further implications of our discussion of SAI.

4. Further Social and Philosophical Implications

Perhaps the best way to introduce these additional issues is to consider that the postbiological cosmos approach involves a shift in our usual perspective about intelligent life in the universe. Normally, we expect that if we encountered advanced alien intelligence we would likely encounter creatures with very different biological features than us. The postbiological cosmos approach suggests that understanding the most advanced intelligences may require that our focus move away from biology to theorizing about the computational abilities of advanced AIs. Further, as we reflect on

[5] For an accessible overview of the debate see Falk (2015).

the nature of postbiological intelligence, we must be keenly aware that we may be reflecting upon the nature of our own *descendants* as well as aliens, as human intelligence may itself become postbiological. In essence, the line between "us" and "them" blurs, and our focus moves away from biology to the tremendously difficult task of understanding the computations and behaviors of creatures that will be far more advanced than we are.

This being said, what would the impact on society be, in the event that we learned that vastly more intelligent beings existed elsewhere in the universe, and that they were not even biological, being SAIs? It would of course depend on various contextual details of the discovery event that are hard to predict, but it seems fair to say that finding out that a superior intelligence had evolved beyond biological life and become synthetic could be rather sobering. In this case, it would be natural for people to ask: Are SAIs, including our own possible postbiological descendants, even selves or persons, or are they just mindless machines? Relatedly, what can we make of the inner lives of such beings? Would it feel a certain way to be them, from the inside? The futurist Ray Kurzweil, who is now a director of engineering at Google, has written extensively of scenarios in which humanity eventually merges with machines. Kurzweil suggests that humans should transcend our biological bodies, reaching a higher level of consciousness, and freeing humans from the confines of biological senescence, and eventually becoming SAI (Kurzweil, 2005). This certainly suggests that SAIs are conscious, and that they are selves, with interests and heightened capacities to appreciate the world. In contrast to Ray Kurzweil's utopian outlook, (an outlook shared by many transhumanists), I do not see normative discussions of the value of post-biological existence in the astrobiology literature on the postbiological cosmos approach. (This is not to say that astrobiological discussions should make normative claims; I am merely making an observation.)

My own view is that the question of whether SAI is conscious is key to how we should value postbiological existence. A SAI could be a sophisticated information processing system, outperforming humans in every cognitive domain, but if it doesn't feel like anything to be an AI, it difficult to view these beings as having the same value as conscious beings, being persons or selves. Consciousness is the philosophical cornerstone here, being a necessary condition on being a self or person, on my view, so it is important to understand what I mean by "consciousness." Consider that every moment of your waking life, and whenever you dream, there is something its *feels* like to be you. When you see the warm hues of a sunset, or hear the scream of an espresso machine, you are having *conscious experience*. Conscious experience includes *all forms of awareness*: e.g., sensory experience, inner thought, and emotion. Bearing this in mind, let us now ask: could an AI be conscious, as Kurzweil and others suggest? I've discussed the issue of consciousness in an earlier astrobiology piece (Schneider 2015), but since this time, (although I still agree with the considerations I raised, especially, the critical discussion of John Searle's earlier case against machine consciousness) two new considerations move me in a more conservative direction, dampening my previous optimism about machine consciousness.

First**,** we know that at least some biological beings can be conscious. For each of us can introspect and tell that we are conscious – right now, you can tell you are experiencing the world. And many of us believe that nonhuman animals are conscious because they are neurophysiologically similar to us. But how do we know something made of computer chips -- a different stuff entirely -- can have experience?

Philosophers often believe that different substrates, when isomorphic in their information processing capacities, will also both function the same when it comes to consciousness (see, e.g., Chalmers 1996). But silicon and carbon differ in important ways, to begin with.[6] First of all, it isn't even clear that they are isomorphic in their information processing abilities, because silicon is faster and more durable than neurons (Bostrom, 2014; Schneider 2105). Second, consider that carbon and silicon differ molecularly. Carbon molecules form stronger, more stable chemical bonds than silicon, which allows carbon to form an extraordinary number of compounds, and unlike silicon, carbon has the capacity to more easily form double-bonds. This difference has important implications in astrobiology, because it is for this reason that carbon, and not silicon, is said to be well-suited for the development of life throughout the universe (Bennett and Shostak, 2012). If these chemical differences impact life itself, we should not rule out the possibility that these chemical differences also impact other key functions, such as whether silicon gives rise to consciousness. This is not a consideration that should alone justify an endorsement of biological naturalism, a view that denies that machines can be conscious, but it is a consideration indicating that it is not yet clear whether AI can be conscious.

If silicon cannot be the basis for consciousness, then superintelligent machines -- machines that may even one day even supplant us -- will exhibit a vastly superior form of intelligence, but they will lack inner experience. Just as the breathtaking android in the movie *Ex Machina* (2015) convinced the main character, Caleb, that she was in love with him, so too, a clever AI may convincingly behave as if it is conscious, but lack consciousness entirely. Further, it would not even matter if a SAI is a BISA, having a roughly similar cognitive architecture as humans, including a global workspace. Although activity in a global workspace is correlated with conscious activity in humans, BISAs made of a substrate that cannot produce consciousness would not have experience.

Yet suppose, for the moment, we find out microchips *are* the right stuff. (Indeed, we propose a test for this in Schneider and Mandik, 2016). A second issue still arises. Even if microchips are the right stuff in principle, it may be more efficient for a superintelligence to *eliminate* consciousness from its processing. For think about how consciousness works in the human case. Only a small percentage of human mental processing is conscious at any given time. And consciousness is correlated with novel learning tasks that require concentration. Consider how focused you were when you first learned to drive, for instance. A superintelligence would surpass expert-level knowledge in every domain, with rapid-fire computations ranging over databases that could include the entire internet and encompass the whole planet. What would be novel to it? What would require slow, deliberative focus? Wouldn't it have mastered everything already?

To find out if a superintelligence is truly conscious, we have to examine the details of the particular SAI's inner organization, and this could be an extraordinary challenge, especially if a SAI is not a BISA or is more advanced than an early SAI. Further, although we may expect certain features in BISAs, such as combinatorial representations,

[6] I focus on silicon as a substrate, but it is important to bear in mind that alternate materials will likely be used. (This is why I often simply say, "microchips", using it as a placeholder expression for whatever is used.) Chips made of carbon nanotubes and graphene are currently under development. Such chips, even if they are roughly "brainlike", will be structured differently than biological cells, and the question of whether they give rise to consciousness arises in these contexts too. Herein, I focus on silicon, as it is in current use.

determining if a superintelligence is conscious would be extremely challenging, because it requires close examination of the machine's architecture in totality. The easiest situation would be encountering an early superintelligence that is a BISA, and we could tell that had a global workspace. But there is still the possibility that a workspace may not even be correlated with consciousness in superintelligences. We may surmise that a SAI has a system that can be functionally decomposed, and that it has combinatorial representations and other features that I've mentioned, but beyond the short window, its design can quickly morph into something too complicated for human understanding (Bostrom 2014). Consciousness likely depends upon the complex interaction of a variety of cognitive and perceptual functions. And how can we recognize which mental processing is conscious when the AIs organization becomes so radically unlike the organization of human and nonhuman animal brains?

In sum, we need to determine when – and even whether – a SAI even needs to be conscious in the first place. Some may, but some may not. And this will be difficult, especially in more complex superintelligences that are not BISAs. It may in fact require an assessment by humans that are themselves highly enhanced intelligences.

The matter of AI consciousness is of significance to whether SAIs are selves or persons, and it will likely be of social concern should we encounter SAIs, whether the SAI be alien or a human creation. Consider, first, a scenario in which humans develop SAI on Earth. As mentioned, some suspect that machines will be the next phase in the evolution of intelligence on Earth. Notice that if it doesn't feel like anything to be an AI, we have to ask whether we want to be a mere intermediate step to AI, even if AI doesn't turn against us, and humanity "merges" with it, as transhumanists envision. In an extreme, horrifying case, humans become postbiological, merging with machines, and only nonhuman animals are left to feel the spark of insight, the pangs of grief, or the warm hues of a sunrise. This would be an unfathomable loss, one that is not offset by a mere net gain in intelligence, and I doubt that this is really what transhumanists like Bostrom and Kurzweil envision for humanity. So, the question of whether AI can be conscious may concern the very future of humanity, and it impacts how we would view superintelligence.[7]

Bearing all this in mind, now consider the possibility of encountering *alien* SAI. It would be natural to ask whether biological intelligence on Earth and throughout the cosmos will be like the case just encountered – that is, whether technological civilizations, in general, evolve toward postbiological existence, as proponents of the postbiological cosmos view suspect. If people suspect so, and if they also suspect that AI isn't conscious, they would likely view the suggestion that intelligence tends to become postbiological with dismay. For even if the universe was stocked full of AIs of unbelievable intelligence, why would nonconscious machines have the same value we place on biological intelligence, which is conscious? Nonconscious machines cannot experience the world -- there is nothing it is like to be them.

So, the issue of machine consciousness is key to how we react to the discovery of SAI. And while I hesitate to speak for world religions, discussions with my colleagues in religious studies and theology at the 2015-2016 NASA funded astrobiology project at the Center of Theological Inquiry, suggest that many would reject the possibility that SAIs have souls, or are somehow made in God's image, if they are not even conscious beings. Pope Francis has recently commented that he would baptise an extraterrestrial

[7] While prematurely judging AI as conscious could be a mistake, so too could judging that AI are non-conscious. Here, the ethical costs are high: assuming them to be non-conscious may cause us to commit grave wrongs against them.

(Consolmagno and Mueller 2014). But I wonder how he would react if asked to baptise SAI, let alone one that is not capable of consciousness.

Additional issues would surely arise as well. For instance, consider an issue from my home discipline, philosophy of mind. Given the variety of possible intelligences, it is an intriguing question to ask whether creatures with different sensory modalities may have the same kind of thoughts or think in a similar ways as humans do. As it happens, there is a debate in the field of philosophy of mind that is relevant to this question. Contemporary neo-empiricists, such as the philosopher Jesse Prinz, have argued that all concepts are modality specific, being couched in a particular sensory format, such as vision (Prinz 2004). If he's correct, it may be difficult to understand the thinking of creatures with vastly different sensory experiences than us. But I am skeptical. For instance, consider my aforementioned comment on viewpoint invariant representations. At a higher level of processing, information seems to become less viewpoint dependent. Similarly, it becomes less modality specific, as with the processing in the human brain, as it ascends from particular sensory modalities to the brain's association areas and into working memory and attention, where it is in a more neutral format.

But these matters are subtle and deserve a lengthier treatment. I pursued issues related to this topic in my monograph, *The Language of Thought*, which looked at whether thinking is independent of the kind of perceptual modalities humans have and is also prior to the kind of language we speak (Schneider 2011b). This view is descended from the groundbreaking work of Jerry Fodor (1978). In the context of SAI, an intriguing question is the following: If there is an inner mental language that is independent of sensory modalities, having the aforementioned combinatorial structure, would this be some sort of intellectual common ground, should we encounter other advanced intelligences? Many of these issues apply to the case of intelligent biological alien life as well, and could also be helpful in the context of the development of SAI on Earth.

5. Conclusion

In this piece, I've discussed why it is likely that the alien civilizations we encounter will be forms of superintelligent AI (or "SAI"). I then turned to the difficult question of how such creatures might think. I provisionally attempted to identify some goals and cognitive capacities likely to be possessed by superintelligent beings. I discuss Nick Bostrom's recent book on superintelligence, which focuses on the genesis of SAI on Earth; as it happens, many of Bostrom's observations were informative in the present context (Bostrom 2014). I then isolated a specific type of superintelligence that is of particular import in the context of alien superintelligence, biologically-inspired superintelligences ("BISAs"). I urged that if any superintelligences we encounter are BISAs, certain work in computational neuroscience, cognitive neuroscience and philosophy of mind may provide resources for at least a rough understanding the computations of BISAs. Finally, I discussed some social implications of encountering superintelligent AI in space or Earth, with special focus on the control problem and the question of whether such beings could be conscious.

References

Baars, B. 2008. "The Global Workspace Theory of Consciousness." In M. Velmans and S. Schneider (eds.), *The Blackwell Companion to Consciousness*. Boston, MA: Wiley-Blackwell, pp. 236-247.

Block, N. 1995. "The Mind as the Software of the Brain." In D. Osherson, L. Gleitman, S. Kosslyn, E. Smith, and S. Sternberg (eds.), *An Invitation to Cognitive Science*. New York: MIT Press, pp. 377–421.

Bostrom N., Chislenko, A., Hughes, J., More, M., Sandberg, A., Vita-More, N., et al,(2003) . "The Transhumanist Frequently Asked Questions": v 2.1. World Transhumanist Association. (The most recent version of this document is reproduced at: http://humanityplus.org/philosophy/transhumanist-faq/)

Bostrom, N. 2005. "History of Transhumanist Thought." *Journal of Evolution and Technology*, 14, 1–25.

Bostrom, N. 2008. "Dignity and Enhancement". In *The President's Council on Bioethics, Human Dignity and Bioethics: Essays Commissioned by the President's Council on Bioethics*, Washington, DC: US Government Printing Office.

Bostrom, N. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bostrom, N and Cirkovic, M., 2008, *Global Catastrophic Risks*, Oxford: Oxford Univ. Press.

Bradbury, R., Cirkovic, M., and Dvorsky, G. 2011. "Dysonian Approach to SETI: A Fruitful Middle Ground?" *Journal of the British Interplanetary Society*, 64, pp. 156–165.

Brin, David. 2015. *Shall We Shout into the Cosmos?* Web. Extracted July 1. 2016. http://www.davidbrin.com/setisearch.html

Chalmers, D. 1996. Absent Qualia, Fading Qualia, Dancing Qualia. In: *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Cirkovic, M. and Bradbury, R. 2006. "Galactic Gradients, Postbiological Evolution and the Apparent Failure of SETI." *New Astronomy* 11, 628–639.

Clarke, A. (1962). *Profiles of the Future: An Inquiry into the Limits of the Possible*. New York, NY: Harper and Row.

Consolmagno, G. and Mueller, P. 2014) ***Would You Baptize an Extraterrestrial?: . . . and Other Questions from the Astronomers' In-box at the Vatican Observatory.*** **Penguin Random House.**

Davies, P. (2010). *The Eerie Science: Renewing Our Search for Alien Intelligence*. Boston: Houghton Mifflin Harcourt.

Dick, S. 2013. "Bringing Culture to Cosmos: the Postbiological Universe." In S. J. Dick, and M. Lupisella (eds.), *Cosmos and Culture: Cultural Evolution in a Cosmic Context*. Washington, DC: NASA, online at http://history.nasa.gov/SP-4802.pdf.

Falk, D. 2015. "Is This Thing On?: The fierce debate over whether we should try to contact extraterrestrial life or wait for aliens to contact us." *Slate*. Web: http://www.slate.com/articles/technology/future_tense/2015/03/active_seti_should_we_reach_out_to_extraterrestrial_life_or_are_aliens_dangerous.html. Extracted July 1, 2016.

Fodor, Jerry (1978). *The Language of Thought*. Boston: MIT Press.

Garreau, J. 2005. *Radical Evolution: The Promise and Peril of Enhancing our Minds, Our Bodies – And What it Means to Be Human*. New York, NY: Doubleday.

Guerini, Federico. 2014. "DARPA's ElectRx Project: Self-Healing Bodies Through Targeted Stimulation Of The Nerves," http://www.forbes.com/sites/federicoguerrini/2014/08/29/darpas-electrx-project-self-healing-bodies-through-targeted-stimulation-of-the-nerves/ *Forbes Magazine*, 8/29/2014. Extracted Sept. 30, 2014.

Madrigal, A. (2013). Is this Virtual Worm the First Sign of the Singularity? *The New Atlantis*, May 17.

Müller, Vincent C. and Bostrom, Nick (forthcoming 2014), "Future progress in artificial intelligence: A Survey of Expert Opinion," in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Synthese Library; Berlin: Springer).

Hawkins, J. and Blakeslee, S. 2004. *On Intelligence: How a New Understanding of the Brain will Lead to the Creation of Truly Intelligent Machine*. NewYork, NY: Times Books.

Holley, Peter. 2015. "Bill Gates on Dangers of Artificial Intelligence: 'I Don't Understand Why Some People Are Not Concerned'." *The Washington Post*. Web. Extracted July 1, 2016.

Kurzweil, R. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York, NY: Viking.

Miller, R. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" *The Psychological Review*, *63*, 81–97.

Prinz, J. 2004. *Furnishing the Mind: Concepts and their Perceptual Basis*. Boston: MIT Press.

Sandberg, A., Boström, N. 2008. "Whole Brain Emulation: A Roadmap." Technical Report #2008•3. Future of Humanity Institute, Oxford University.

Schneider, S. 2011a. “Mindscan: Transcending and Enhancing the Brain.” In J. Giordano (ed.), *Neuroscience and Neuroethics: Issues At the Intersection of Mind, Meanings and Morality*. Cambridge: Cambridge University Press.

Schneider, S. 2011b. *The Language of Thought: a New Philosophical Direction*. Boston, MA: MIT Press.

Schneider, S. 2014. “The Philosophy of ‘Her.’” *The New York Times*, March 2.

Schneider, S. 2015. “Alien Minds.” In *Discovery*, Steven Dick, ed., Cambridge, Cambridge University Press.

Schneider and Mandik, “The Future of Philosophy of Mind,” Amy Kind (ed)., forthcoming.

Seung, S. 2012. *Connectome: How the Brain’s Wiring Makes Us Who We Are*. Boston, MA: Houghton Mifflin Harcourt

Shostak, S. 2009. *Confessions of an Alien Hunter*. New York, NY: National Geographic.

Shostak, S. 2015. “Should We Keep a Low Profile in Space?” *The New York Times*. Web. Extracted July 1, 2016.

Wallach, W. and Allen, C. (2008) *Moral Machines*. Oxford: OUP.

Yudkowsky, E. 2008. “Artificial Intelligence as a Positive and Negative Factor in Global Risk.” In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan. M. Ćirković. Oxford University Press. Pp. 308-345.