12 Alien minds

SUSAN SCHNEIDER

How would intelligent aliens think? Would they have conscious experiences? Would it feel a certain way to be an alien? It is easy to dismiss these questions as too speculative, since we haven't encountered aliens, at least as far as we know. And in conceiving of alien minds we do so from *within* – from inside the vantage point of the sensory experiences and thinking patterns characteristic of our species. At best, we anthropomorphize; at worst, we risk stupendous failures of the imagination.

Still, ignoring these questions could be a grave mistake. Some proponents of the search for extraterrestrial intelligence (SETI) estimate that we will encounter alien intelligence within the next several decades. Even if you hold a more conservative estimate – say, that the chance of encountering alien intelligence in the next 50 years is 5 percent – the stakes for our species are high. Knowing that we are not alone in the universe would be a profound realization, and contact with an alien civilization could produce amazing technological innovations and cultural insights. It thus can be valuable to consider these questions, albeit with the goal of introducing possible routes to answering them, rather than producing definitive answers. So, let us ask: how might aliens think? And, would they be conscious? Believe it or not, we can say something concrete in response to both of these questions, drawing from work in philosophy and cognitive science.

You might think the second question is odd. After all, if aliens have sophisticated enough mental lives to be intelligent, wouldn't they be conscious? The far more intriguing question is: what would the quality of their consciousness be like? This would be putting the cart before the horse, however, since I do not believe that most advanced alien civilizations will be biological. The most sophisticated civilizations will be postbiological, forms of artificial intelligence (AI). (Cirkovic and Bradbury 2006; Shostak 2009; Davies 2010, 153–168; Bradbury *et al.* 2011; Dick 2013).[1] Further, alien

[1] "Postbiological," in the astrobiology literature contrasts with "posthuman" in the singularity literature. In the astrobiology literature "postbiological" creatures are forms of AI. In the singularity literature "posthumans" can be forms of AI, but they need not be. They are merely creatures who are descended from humans but which have alterations that make them no longer unambiguously human. They need not be full-fledged AI.

civilizations will tend to be forms of *superintelligence*: intelligence that is able to exceed the best human-level intelligence in every field – social skills, general wisdom, scientific creativity, and so on (Kurzweil 2005, Schneider 2011a, Bostrom 2014). It is a substantive question whether superintelligent AI (SAI) could have conscious experiences; philosophers have vigorously debated just this question of in the case of AI in general. Perhaps all their information processing happens in the dark, so to speak, without any inner experience at all. This is why I find the second question so pressing, and in an important sense prior to any inquiry as to the contours of alien consciousness, and prior to the epistemological problem of how we can know "what it is like" to be an alien.

In this chapter I first explain why it is likely that the alien civilizations we encounter will be forms of SAI. I then turn to the question of whether superintelligent aliens can be conscious – whether it feels a certain way to be an alien, despite their non-biological nature. Here, I draw from the literature in philosophy of AI, and urge that although we cannot be *certain* that superintelligent aliens can be conscious, it is likely that they would be. I then turn to the difficult question of how such creatures might think. I provisionally attempt 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' observations are informative in the present context. Finally, I isolate a specific type of superintelligence that is of particular import in the context of alien superintelligence, biologically inspired superintelligences ("BISAs").

Alien superintelligence

SETI programs have been searching for biological life. Our culture has long depicted aliens as humanoid creatures with small, pointy chins, massive eyes, and large heads, apparently to house brains that are larger than ours. Paradigmatically, they are "little green men." While we are aware that our culture is anthropomorphizing, I imagine that my suggestion that aliens are supercomputers may strike you as far-fetched. So what is my rationale for the view that most intelligent alien civilizations will have members that are forms of SAI? I offer three observations that, together, motivate this conclusion.

(1) The short window observation. Once a society creates the technology that could put them in touch with the cosmos, they are only a few hundred years away from changing their own paradigm from biology to AI. (Shostak 2009; Davies 2010, 153–168; Dick 2013,). This "short window" makes it more likely that the aliens we encounter would be postbiological.

The short-window observation is supported by human cultural evolution, at least thus far. Our first radio signals date back only about 120 years, and space exploration is only about 50 years old, but we are already immersed in digital technology, such as cell-phones and laptop computers. Devices such as the Google Glass promise to bring the Internet into more direct contact with our bodies, and it is probably a matter of less than 50 years before sophisticated internet connections are wired directly into our brains. Indeed, implants for Parkinson's are already in use, and in the United States the Defense Advanced Research Projects Agency (DARPA) has started to develop neural implants that interface directly with the nervous system, regulating conditions such as post-traumatic stress disorder, arthritis, depression, and Crohn's disease. DARPA's program, called "ElectRx," aims to replace certain medications with "closed-loop" neural implants, implants that continually assess the state of one's health, and provide the necessary nerve stimulation to keep one's biological systems functioning properly (Guerini, 2014). Eventually, implants will be developed to enhance normal brain functioning, rather than for medical purposes.

Where might all this all lead? A thought experiment from my "Transcending and Enhancing the Human Brain" is suggestive (Schneider, 2011a).

> Suppose it is 2025 and being a technophile, you purchase brain enhancements as they become readily available. First, you add a mobile internet connection to your retina, then, you enhance your working memory by adding neural circuitry. You are now officially a cyborg. Now skip ahead to 2040. Through nanotechnological therapies and enhancements you are able to extend your lifespan, and as the years progress, you continue to accumulate more far-reaching enhancements. By 2060, after several small but cumulatively profound alterations, you are a "posthuman." To quote philosopher Nick Bostrom, posthumans are possible future beings, "whose basic capacities so radically exceed those of present humans as to be no longer unambiguously human by our current standards" (Bostrom 2003).
>
> At this point, your intelligence is enhanced not just in terms of speed of mental processing; you are now able to make rich connections that you were not able to make before. Unenhanced humans, or "naturals," seem to you to be intellectually disabled – you have little in common with them – but as a transhumanist, you are supportive of their right to not enhance (Bostrom 2003; Garreau 2005; Kurzweil 2005).
>
> It is now AD 2400. For years, worldwide technological developments, including your own enhancements, have been facilitated by superintelligent AI. . . . Indeed, as Bostrom explains, "creating superintelligence may be the last invention that humans will ever need to make, since superintelligences could themselves take care of further

scientific and technological developments" (Bostrom *et al.* 2003). Over time, the slow addition of better and better neural circuitry has left no real intellectual difference in kind between you and superintelligent AI. The only real difference between you and an AI creature of standard design is one of origin – you were once a natural. But you are now almost entirely engineered by technology – you are perhaps more aptly characterized as a member of a rather heterogeneous class of AI life forms (Kurzweil 2005).

Of course, this is just a thought experiment. But I've just observed that we are already beginning to develop neural implants. It is hard to imagine people in mainstream society resisting opportunities for superior health, intelligence, and efficiency. And just as people have already turned to cryonics, even in its embryonic state, I suspect that they will increasingly try to upload to avoid death, especially as the technology is perfected.[2] Indeed, the Future of Humanity Institute at Oxford University (Sandberg and Boström 2008) has released a report on the technological requirements for uploading a mind to a machine. And a Defense Department agency has funded a program, *Synapse*, which is developing a computer that resembles a brain in form and function (Schneider 2014). In essence, the short-window observation is supported by our own cultural evolution, at least thus far.

You may object that this argument employs "$N = 1$ reasoning," generalizing from the human case to the case of alien civilizations (see Chapter 7 in this volume). Still, it is 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 civilizations will develop technologies to advance their intelligence and survival. And, as I will explain in a moment, silicon is a better medium for thinking than carbon.

A second objection to my short-window observation rightly points out that nothing I have said thus far suggests that humans will be *superintelligent*. I have merely said that future humans will be *posthuman*. While I offer support for the view that our own cultural evolution suggests that humans will be post-biological, this does not show that advanced alien civilizations will reach superintelligence. So even if one is comfortable reasoning from the human case, the human case does not support the position that the members of advanced alien civilizations will be superintelligent.

This is correct. This is the task of the second observation.

(2) The greater age of alien civilizations. Proponents of SETI have often concluded that alien civilizations would be much older than our own: ". . . all lines of evidence converge on the conclusion that the maximum age of

[2] Although I have elsewhere argued that uploading would merely create a copy of one's brain configuration and would not be a true means of survival, I doubt dying individuals will act on a philosopher's qualms when they have little to lose by trying (Schneider 2014).

extraterrestrial intelligence would be billions of years, specifically [it] ranges from 1.7 billion to 8 billion years" (Dick 2013, 468). If civilizations are millions or billions of years older than us, many would be vastly more intelligent than we are. By our standards, many would be superintelligent. We are galactic babies.

But would they be forms of AI, as well as forms of superintelligence? I believe so. Even if they were 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 carbon-based, as silicon is a better medium for intelligence. I expect that they will not be carbon-based. Uploading allows a creature near immortality, enables reboots, and allows it to survive under a variety of conditions that carbon-based life forms cannot. In addition, silicon appears to be a better medium for information processing than the brain itself. Neurons reach a peak speed of about 200 Hz, which is seven orders of magnitude slower than current microprocessors (Bostrom 2014, 59). While the brain can compensate for some of this with massive parallelism, features such as "hubs," and so on, crucial mental capacities, such as attention, rely upon serial processing, which is incredibly slow, and has a maximum capacity of about seven manageable chunks (Miller 1956). Further, the number of neurons in a human brain is limited by cranial volume and metabolism, but computers can occupy entire buildings or cities, and can even be remotely connected across the globe (Bostrom 2014). Of course, the human brain is far more intelligent than any modern computer. But intelligent machines 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 forms of SAI, because silicon is a superior medium for superintelligence. From this I conclude that many advanced alien civilizations will be populated by forms of SAI.

Even if I am wrong – 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 form of SAI. 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.

All this being said, would superintelligent aliens be conscious, having inner experiences? Here, I draw from a rich philosophical literature on the nature of conscious experience.

Would superintelligent aliens be conscious?

Consider your own conscious experience. Suppose that you are sitting in a cafe preparing to give a lecture. All in one moment, you taste the espresso you sip, consider an idea, and hear the scream of the espresso machine. This is your current stream of consciousness. Conscious streams seem to be very much bound up with who you are. It is not that *this* particular moment is essential – although you may feel that certain ones are important. It is rather that throughout your waking life, you seem to be the subject of a unified stream of experience that presents you as the subject, viewing the show.

Let us focus on three features of the stream: first, it may seem to you, put metaphorically, that there is a sort of “screen” or “stage” in which experiences present themselves to your “mind’s eye.” That is, there appears to be a central place where experiences are “screened” before you. Daniel Dennett calls this place “the Cartesian Theater” (Dennett 1991). Second, in this central place there seems to be a singular point in time which, given a particular sensory input, consciousness happens. For instance, there seems to be one moment in which the scream of the espresso machine begins, pulling you out of your concentration. Finally, there appears to be a self – someone who is inside the theater, watching the show.

Philosophers have considered each of these features in detail. Each is highly problematic. For instance, an explanation of consciousness cannot literally be that there is a mind’s eye in the brain, watching a show. And there is no evidence that there is a singular place or time in the brain where consciousness congeals.

These are intriguing issues, but pursuing them in the context of alien consciousness is putting the cart before the horse. For there is a more fundamental problem: would superintelligent aliens, being forms of AI, even be conscious? Why should we believe that creatures so vastly different from us, being silicon-based, would have inner experience at all?

This problem relates to what philosophers call the *hard problem of consciousness*, a problem that was posed in the context of human consciousness by the philosopher David Chalmers (Chalmers 2008). Chalmers’ hard problem is the following. As cognitive science underscores, when we deliberate, hear music, see the rich hues of a sunset, and so on, there is information processing going on in the brain. But above and beyond the manipulation of data, there is a subjective side – there is a "felt quality" to our experience. The hard problem asks: why does all this information processing in the human brain, under certain conditions, have a felt quality to it?

As Chalmers emphasizes, the hard problem is a philosophers' problem, because it doesn't seem to have a scientific answer. For instance, we could develop a complete theory of vision, understanding all of the details of visual processing in the brain, but still not understand why there are subjective experiences attached to these informational states. Chalmers contrasts the hard problem with what he calls "easy problems," problems involving consciousness that have eventual scientific answers, such as the mechanisms behind attention and how we categorize and react to stimuli. Of course these scientific problems are difficult problems; Chalmers merely calls them "easy problems" to contrast them with the "hard problem" of consciousness, which he thinks will not have a purely scientific solution.

We now face yet another perplexing issue involving consciousness – a kind of "hard problem" involving alien superintelligence, if you will: *the hard problem of alien superintelligence*. Would the processing of a silicon-based superintelligent system feel a certain way, from the inside? An alien SAI could solve problems that even the brightest humans are unable to solve, but still, being made of a non-biological substrate, would its information processing feel a certain way from the inside?

It is worth underscoring that the hard problem of superintelligence is not just Chalmers' hard problem of consciousness applied to the case of aliens. For the hard problem of consciousness assumes that we are conscious – after all, each of us can tell from introspecting that we are conscious at this moment. It asks *why* we are conscious. Why does all your information processing feel a certain way from the inside? In contrast, the hard problem of alien consciousness asks *whether* alien superintelligence, being silicon-based, is even capable of being conscious. It does not presuppose that alien superintelligence is conscious. These are different problems, but they are both hard problems that science alone cannot answer.

The problem in the case of superintelligent aliens is that the capacity to be conscious may be unique to biological, carbon-based, organisms. According to *biological naturalism* even the most sophisticated forms of AI will be devoid of inner experience (Searle 1980, Blackmore 2004, Searle 2008). Indeed, even humans wishing to upload their minds will fail to transfer their consciousness. Although they may copy their memories onto a computational format, their consciousness will not transfer, since biological naturalists hold that consciousness requires a biological substrate.[3]

What arguments support biological naturalism? The most common consideration in favor of biological naturalism is John Searle's Chinese Room thought experiment, which is said to suggest that a computer program cannot understand or be conscious (Searle 1980). Searle supposes that he's locked in a room, where he's handed a set of English rules that allow him to link one set of Chinese symbols with other Chinese symbols. So although he doesn't know Chinese, the rules allow him to respond, in written Chinese, to questions written in Chinese. So he is essentially processing symbols. Searle concludes that although those outside of the room may think he understands Chinese, he obviously doesn't; similarly, a computer may appear to be having a Chinese conversation, yet it does not truly understand Chinese. Nor is it conscious.

Although it is correct that Searle doesn't understand Chinese, the issue is not really whether Searle understands; Searle is just one part of the larger system. The relevant question is whether *the system as a whole* understands Chinese. This basic response to Searle's Chinese Room thought experiment is known as the Systems Reply.[4]

It strikes me as implausible that a simple system like the Chinese Room understands, however, for the Chinese Room is not complex enough to understand or be conscious. But the Systems Reply is onto something: the real issue is whether the *system* as a whole understands, not whether one component does. This leaves open the possibility that a more complex silicon-based system could understand; of course, the computations of a superintelligent AI will be far more complex than the human brain.

Here, some might suspect that we could just reformulate the Chinese Room thought experiment in the context of an SAI. But what is fueling this suspicion? It cannot be that some central component in the SAI, analogous to Searle in the Chinese Room, doesn't understand, for we've just observed that it is the system as a whole that understands. Is the suspicion instead fueled by the position that

[3] Biological naturalism was originally developed by John Searle, who developed the view in the context of a larger account of the relation between the mind and body. I will not discuss these details, and they are not essential to the position I've just sketched. Indeed, it isn't clear that Searle is still a biological naturalist, although he persists in calling his view "biological naturalism." In his chapter to my recent *Blackwell Companion to Consciousness* he wrote: "The fact that brain processes cause consciousness does not imply that only brains can be conscious. The brain is a biological machine, and we might build an artificial machine that was conscious; just as the heart is a machine, and we have built artificial hearts. Because we do not know exactly how the brain does it we are not yet in a position to know how to do it artificially." (Searle 2008)

[4] For a thorough treatment of the responses to Searle's argument, including the system's reply, the reader may turn to the comments appearing with Searle's original piece, Searle (1980) as well as Cole (2014).

understanding and consciousness do not decompose into more basic operations? If so, then the thought experiment purports to prove too much. Consider the case of the human brain. According to cognitive science, cognitive and perceptual capacities decompose into more basic operations, which are themselves decomposable into more basic constituents, which themselves can be explained causally (Block 1995). If the Chinese Room illustrates that mentality cannot be explained like this, then the brain cannot be explained in this manner either. But this explanatory approach, known as "the method of functional decomposition," is a leading approach to explaining mental capacities in cognitive science. Consciousness and understanding are complex mental properties that are determined by the arrangements of neurons in the brain.

Further, biological naturalism denies one of the main insights of cognitive science – the insight that the brain is computational – without substantial empirical rationale. Cognitive science suggests that our best empirical theory of the brain holds that the mind is an information processing system and that all mental functions are computations. If cognitive science is correct that thinking is computational, then humans and SAI share a common feature: their thinking is essentially computational. Just as a phone call and a smoke signal can convey the same information, thought can have both silicon- and carbon-based substrates. The upshot is that if cognitive science is correct that thinking is computational, we can also expect that sophisticated thinking machines can be conscious, although the contours of their conscious experiences will surely differ.

Indeed, I've noted that silicon is arguably a better medium for information processing than the brain. So why isn't silicon a *better* medium for consciousness, rather than a *worse* one, as the biological naturalists propose? It would be surprising if SAI, which would have far superior information processing abilities than we do, turned out to be deficient with respect to consciousness. For our best scientific theories of consciousness hold that consciousness is closely related to information processing (Baars 2008, Tonini 2008).

Some would point out that to show that AI cannot be conscious, the biological naturalist would need to locate a special consciousness property (call it "P"), which inheres in neurons or their configurations, and which cannot be instantiated by silicon. Thus far, P has not been discovered. It isn't clear, however, that locating P would prove biological naturalism to be correct. For the computationalist can just say that machines are capable of instantiating a different type of consciousness property, F, which is specific to silicon-based systems.

Massimo Pigliucci has offered a different kind of consideration in favor of biological naturalism, however. He sees philosophers who argue for computationalism as embracing an implausible perspective on the nature of consciousness: functionalism. According to *functionalists* the nature of a mental state depends on the way it functions, or the role it plays in the system of which it is a part. Pigliucci is correct that traditional functionalists, such as Jerry Fodor, generally mistakenly ignore the biological workings of the brain. Pigliucci objects: ". . . functionality isn't just a result of the proper arrangement of the parts of a system, but also of the types of materials (and their properties) that make up those parts" (Pigliucci 2014).

Fodor's well-known antipathy towards neuroscience should not mislead us into thinking that functionalism must ignore neuroscience, however. Clearly, any well-conceived functionalist position must take into consideration neuroscientific work on the brain because the functionalist is interested in the causal or dispositional properties of the parts, not just the parts themselves. Indeed, as I've argued in my book *The Language of Thought*, viewing the brain as irrelevant to the computational approach to the mind is a huge mistake. The brain is the best computational system we know of (Schneider 2011b).

Does this make my position a form of biological naturalism? Not in the least. I am suggesting that viewing neuroscience (and by extension, biology) as being opposed to computationalism is mistaken. Indeed, neuroscience is computational; a large subfield of neuroscience is called "computational neuroscience," and it seeks to understand the sense in which the brain is computational and to provide computational accounts of mental capacities identified by related subfields, such as cognitive neuroscience. What makes my view different from biological naturalism is that I hold that thinking is computational, and further, that at least one other substrate besides carbon (i.e. silicon) can give rise to consciousness and understanding, at least in principle.

But biological naturalism is well worth considering. I am reasoning that a substrate that supports superintelligence, being capable of even more sophisticated informational processing than we are, would likely also be one that is conscious. But notice that I've used the expression "likely." For we can never be *certain* that AI is conscious, even if we could study it up close. The problem is akin to the philosophical puzzle known as *the problem of other minds* (Schneider 2014). The problem of other minds is that although you can know that you are conscious, you cannot be certain that other people are conscious as well. After all, you might be witnessing behavior with no accompanying conscious component. In the face of the problem of other minds, all you can do is note that other people have brains that are structurally similar to your own and conclude that since you yourself are conscious, others are likely to be as well. When confronted with AI your predicament would be similar, at least if you accept that thinking is computational. While we couldn't be absolutely certain that an AI program genuinely felt anything, we can't be certain that other humans do either. But it would seem probable in both cases.

So, to the question of whether alien superintelligence can be conscious, I answer, very cautiously, "probably."

How might superintelligent aliens think?

Thus far, I've said little about the structure of superintelligent alien minds. And little is all we can say: superintelligence is by definition a kind of intelligence that outthinks humans in every domain. In an important sense, we cannot predict or fully understand how it will think. Still, we may be able to identify a few important characteristics, albeit in broad strokes.

Nick Bostrom's recent book on superintelligence focuses on the development of superintelligence on Earth, but we can draw from his thoughtful discussion (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's suggests (Bostrom 2014, 107):

> *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 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 may initially strike you as harmless endeavor, 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. Indeed, Bostrom warns that superintelligence emerging on Earth could be of an unpredictable nature, being "extremely alien” to us (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 whatsoever. 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. “uploading”).[5]

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 (Bostrom 2014, 109):

> *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 instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents.

The goals that he 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.

BISAs are of particular interest in the context of alien superintelligence. For if Bostrom is correct that there are many ways 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 out there*. 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 modeled 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

[5] 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.

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.

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:

(1) BISAs are descended from creatures that had motivations like: 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 common to members of many superintelligent alien civilizations? I suspect so.

Consider (1). Intelligent biological life tends to be primarily concerned with its own 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 that, given the massive amounts of computational resources at their disposal, BISAs would 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.

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 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.

If (2) is correct, we might expect the following, for instance.

(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 withoutviewpoint 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. 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 the 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 not having a program consisting of causally interrelated elements each of which consists in causally organized elements.

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. It may be that any truly advanced civilization will have technologies that will be indistinguishable from magic, as Arthur C. Clarke suggested (1962). I obviously 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.

Conclusion

I have argued that the members of the most advanced alien civilizations will be forms of superintelligent artificial intelligence (SAI). I have further suggested, very provisionally, that we might expect that if a given alien superintelligence is a biologically-inspired superintelligent alien (BISA), it would have combinatorial representations and that we could seek insight into its processing by decomposing its computational functions into causally interacting parts. We could also learn about it by looking at the brain wiring diagrams (connectomes) of the members of the original species. Further, BISAs may have one or more global workspaces. Furthermore, I have argued that there is no reason in principle to deny that SAIs could have conscious experience.

Acknowledgements

Many thanks to Joe Corabi, Steven Dick, Clay Ferris Naff, and Eric Schwitzgebel for helpful written comments on an earlier draft and to James Hughes for helpful conversation.

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.

Blackmore, S. 2004. *Consciousness: An Introduction*. New York, NY: Oxford University Press.

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., 2003. "The Transhumanist Frequently Asked Questions": v 2.1. World Transhumanist Association. Retrieved from http://humanityplus.org/philosophy/transhumanist-faq/.

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

Chalmers, D. 2008. "The Hard Problem of Consciousness." In M. Velmans and S. Schneider, *The Blackwell Companion to Consciousness*. Boston, MA: Wiley-Blackwell, 225–236.

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.

Cole, D. 2014. "The Chinese Room Argument", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), E. N. Zalta (ed.), online at http://plato.stanford.edu/archives/sum2014/entries/chinese-room/.

Davies, P. 2010. *The Eerie Silence: Renewing Our Search for Alien Intelligence*. Boston, MA: Houghton, Mifflin, Harcourt.

Dennett, D. 1991. *Consciousness Explained*. New York, NY: Penguin Press.

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.

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," online at http://www.forbes.com/sites/federicoguerrini/2014/08/29/darpas-electrx-project-self-healing-bodies-through-targeted-stimulation-of-the-nerves/.

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.

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

Pigliucci, M. 2014. "Mind Uploading: A Philosophical Counter-Analysis." In R. Blackford and D. Broderick (eds.), *Intelligence Unbound: The Future of Uploaded and Machine Minds*. Boston, MA: Wiley-Blackwell, pp. 119–130.

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.

Searle, J. 1980. “Minds, Brains and Programs.” *The Behavioral and Brain Sciences*, 3, 417–457.

Searle, J. 2008. “Biological Naturalism.” In M. Velmans and S. Schneider (eds.), *The Blackwell Companion to Consciousness*. Boston, MA: Wiley-Blackwell.

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.

Tonini, G. 2008. “The Information Integration Theory of Consciousness.” In M. Velmans and S. Schneider (eds.), *The Blackwell Companion to Consciousness*. Boston, MA: Wiley-Blackwell, pp. 287–300.