12

HOW PHILOSOPHY OF MIND CAN SHAPE THE FUTURE

Susan Schneider and Pete Mandik

A bright metallic thread of future-oriented thinking runs through the tapestry of the philosophy of mind, especially in those parts of the field that have grappled with the possibility of minds as machines. Can a robot feel pain? Can a suitably programmed computer think actual thoughts? Could humans survive the total replacement of their nervous system by neural prosthetics? As the pace of technological change quickens, what was once purely speculative is becoming more and more real. As society moves further into the 21st century, what are the ways that philosophy of mind can shape the future? What challenges will the future bring to the discipline? In this chapter, we examine a few suggestive possibilities. We begin with what we suspect will be a game changer – the development of AI and artificial general intelligence (AGI). We then turn to radical brain enhancements, urging that the future will likely introduce exciting new issues involving (inter alia) the extended mind hypothesis, the epistemology of evaluating the thoughts of vastly smarter beings, mind uploading, and more.

1. The rise of the machines: some philosophical challenges

These last few years have been marked by the widespread cultural recognition that sophisticated AI is under development, and may change the face of society. For instance, according to a recent 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-percent probability by the year 2024. Further, they assign a 50-percent probability by 2050, and a 90-percent probability by 2070 (Muller and Bostrom 2014).[1] AI critics, such as John Searle, Jerry Fodor and Hubert Dreyfus, must now answer to the impressive work coming out of venues like Google's *DeepMind* and exhibited by IBM's *Watson* program,[2] rather than referring back to the notorious litany of failures of AI in the 1970s and 1980s.

Indeed, silicon seems to be a better medium for information processing than the brain. Neurons reach a peak speed of about 200 Hz, which is about seven orders of magnitude slower than current microprocessors (Bostrom 2014, 59). Although the brain compensates 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; Schneider 2014). Additionally, 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; Schneider 2014).

Of course, the human brain is more intelligent than any modern day computer. Intelligent machines can in principle be constructed by reverse engineering the brain, however, 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 different locations at once, is easily modifiable, and can survive under a variety of conditions that carbon-based life cannot. The increases in redundancy and backups that programs allow mean that AI minds will be hardier and more reliable than their biological counterparts.

We've noted AI experts' projections that sophisticated AI may be reached within the next several decades. By "sophisticated AI" what is meant is **artificial general intelligence** (AGI). An AGI is a flexible, domain-general intelligence – an intelligence that can integrate material from various domains, rather than merely excelling at a single task, like winning *Jeopardy* or playing chess. Philosophers have debated the possibility of AGI for decades, and we hope they will help shape the global understanding of AGI in the future. For instance, perhaps some philosophers will discover a distinctively philosophical reason for believing that, despite the successes of Watson and *DeepMind*, experts will (and must) hit a wall when it comes to creating AGI – perhaps computers can excel at domain specific reasoning but general purpose reasoning is not amenable to computational explanation. Or perhaps the resources of the philosophy of mind will not unearth a deep obstacle to AGI, but instead provide insights that will aid in its development.

In any case, within society at large, the earlier skepticism about AGI has given way. Indeed, there is now a general suspicion that once AGI is reached, it may upgrade itself to even greater levels of intelligence. As David Chalmers explains:

> The key idea is that a machine that is more intelligent than humans will be better than humans at designing machines. So it will be capable of designing a machine more intelligent than the most intelligent machine that humans can design. So if it is itself designed by humans, it will be capable of designing a machine more intelligent than itself. By similar reasoning, this next machine will also be capable of designing a machine more intelligent than itself. If every machine in turn does what it is capable of, we should expect a sequence of ever more intelligent machines.
>
> (Chalmers 2010)

In a similar vein, Nick Bostrom's *New York Times* bestselling book *Superintelligence: Paths, Dangers and Strategies* (2014) argues that a superintelligence could

supplant humans as the dominant intelligence on the planet, and that the sequence of changes could be rapid-fire (see also Kurzweil 2005). Indeed, due in large part to Bostrom's book, and the successes at *DeepMind*, this last year marked the widespread cultural and scientific recognition of the possibility of "superintelligent AI."[3]

> **Superintelligent AI**: a kind of artificial general intelligence that is able to exceed the best human level intelligence in every field – social skills, general wisdom, scientific creativity, and so on
> (Bostrom 2014; Kurzweil 2005; Schneider 2009a; 2015).

Superintelligent AI (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, Bostrom warns that 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. Further, a clever machine could bypass safeguards like kill switches and attempts to box it in, and could potentially be an existential threat to humanity (Bostrom 2014). A superintelligence is, after all, defined as an entity that is more intelligent than humans, in every domain. Bostrom calls this problem "The Control Problem." (Bostrom 2014)

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, commented that superintelligent AI could threaten the human race, having goals that humans can neither predict nor control. Yet 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 2010).

The possibility of human or beyond-human AI raises further philosophical questions as well. If AGI and SAI are developed, would they be conscious? Would they be selves or persons, although they are arguably not even living beings? Of course, perhaps we are putting the cart before the horse in assuming that superintelligence can even be developed: perhaps the move from human-level AGI to superintelligence is itself questionable (Chalmers 2010)? After all, how can humans create beyond-human intelligence given that our own intellectual resources are only at a human level? Quicker processing speed and a greater number of cognitive operations do not necessarily result in a qualitative shift to a greater form of intelligence. Indeed, what are markers for "beyond human intelligence", and how can we determine when it has been reached?

In his groundbreaking paper on the singularity, Chalmers suggests even more issues that philosophers could explore:

> Philosophically: The singularity raises many important philosophical questions. . . . The potential consequences of an intelligence explosion force us to think hard about values and morality and about consciousness and personal identity. In effect, the singularity brings up some of the hardest traditional questions in philosophy and raises some new philosophical questions as well.
>
> . . . To determine whether an intelligence explosion will be a good or a bad thing, we need to think about the relationship between intelligence and value. To determine whether we can play a significant role in a post-singularity world, we need to know whether human identity can survive the enhancing of our cognitive systems, perhaps through uploading onto new technology. These are life-or-death questions that may confront us in coming decades or centuries. To have any hope of answering them, we need to think clearly about the philosophical issues.
>
> (Chalmers 2010)

What sorts of things can philosophers do to help tackle the issues raised by AI, the singularity, and other technologies on the horizon? We recommend an approach that draws on thought experiments of the sort traditionally considered by philosophers of mind, but tempered by knowledge of contemporary advances in science and technology.

Philosophers often view thought experiments as windows into the fundamental nature of things – hypothetical situations in the "laboratory of the mind" that depict something that exceeds the bounds of current technology or even is incompatible with the laws of nature, but that is supposed to reveal something
AuQ7 philosophically enlightening about the topic in question (Schneider 2009). Thought experiments can entertain, illustrate a puzzle, lay bare a contradiction in thought, and move us toward further clarification. Yet experimental philosophers have countered that thought experiments are not trustworthy guides to philosophical issues because they covertly rely upon intuitive judgments about possibility that are hostage to features like our cultural and economic backgrounds.

Emerging technologies introduce a host of real world cases – cases that seem nomologically and technologically possible – rather than relying upon dubious intuitions about what is possible in remote possible worlds like zombie worlds (worlds in which no entity is conscious, even the entities that act like they are) or Cartesian worlds stocked with disembodied minds. And, in the domain of emerging technologies – this arena in which science fiction meets science fact – philosophy quite possibly becomes a matter of life and death, as we will further discuss
AuQ8 shortly (Chalmers 2010; Schneider 2009; Mandik 2015).

In what follows, we identify more ways that philosophers of mind can help shape the 21st century. We begin with a fictional scenario that introduces issues about the extended mind hypothesis. We then turn to several interrelated philosophical problems, based upon this scenario and others that we introduce.

2. The ethics of brain enhancement, the extended mind, and human integration into a post-singularity world

Consider the following thought experiment, modified from Schneider (2009a):

> 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 2017)

> 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 (a proponent of the sorts of cybernetic and genetic modifications that, in the extreme case, leads to posthumans), you are supportive of their right to not enhance.
>
> (Bostrom 2017; Garreau 2004; Kurzweil 2005)

> It is now 2250 AD. Over time, the slow addition of better and better neural circuitry has left no real intellectual difference in kind between you and AI. Your mental operations have been gradually transferring to the cloud, and by this point, you are silicon-based. 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.

Of course, this is just a thought experiment, but it is hard to imagine people in mainstream society resisting opportunities for superior health, intelligence,

extreme longevity and efficiency. Indeed, the advanced technologies wing of the defense department (DARPA) is now working on brain chips, electronic prosthetics implanted in the brain, providing intriguing examples of "cyborgs."

There are many philosophical issues that this thought experiment raises. Let us consider a few.

2.1 The extended mind, 2.0

Despite being implanted *in* brains, brain chips strike us as providing better support for the extended mind hypothesis than Clark and Chalmers's original examples of laptops and notepads. (The extended mind hypothesis is the proposal that the physical substrate for the human mind is not restricted to the human central nervous system, but can sometimes or perhaps always include external physical items, as when one's memories are stored in external media such as notebooks and hard drives. See Chapter 10). For it can be objected that laptops and notepads do not seem to exhibit a sufficiently rich cognitive integration with the brain to justify the claim that the mind is extended beyond the brain. Instead, information from notebooks and laptops enters into cognitive and perceptual systems through sensory transducers. When one forgets their laptop or notebook, they only have recourse to the processing of their brain. The brain itself seems to be the true unit of mentality. In contrast, brain implants could become well-integrated with the biological brain, for the inputs from the implants do not enter the cognitive system through sensory transducers, but could in principle function like actual minicolumns or brain regions.

You might object that it is unclear what's "extended" about neural prostheses. If they aren't outside of the body, how do they make the mind "extended"? But if one believes the mind is just the brain, then this makes the mind extended. Further, these implants need not be in the skull, they could be located elsewhere in the body, or even on the cloud, for instance. What is crucial is that they are as well integrated as components of the brain normally are. Would brain or cloud-based implants provide better support for the view that the mind is extended? Further, can consciousness (as opposed to mere information processing) really extend beyond the biological brain? That is, can silicon minicolumns or microchips be part of the neural basis of conscious experience? These are issues well-worth considering, we believe, as we move to a future with neural enhancements and therapies that extend beyond the biological brain.

2.2 Human integration into a post-singularity world

Let us continue our thought experiments further into the future. Suppose that it is now AD 2250 and some humans have upgraded to become superintelligent beings, through gradual cognitive enhancements, including cloud-based computations. But suppose you resist any upgrades – you opt to stay a "Natural" – a member of a group resisting enhancements (Garreau 2004). Having conceptual

resources beyond your wildest imagination, the superintelligent beings generate an entirely new budget of solutions to longstanding, central philosophical problems, such as the mind-body problem, the hard problem of consciousness, and the problem of free will. They univocally and passionately tell you that the solutions are obvious. But you and the other Naturals throw your hands up; these "solutions" strike you and the other unenhanced as gibberish (Schneider 2009b).

You think: Who knows, maybe these "superintelligent" beings were engineered poorly; or maybe it is me. Perhaps the unenhanced are "cognitively closed," as Colin McGinn has argued, being constitutionally unable to solve major philosophical problems (McGinn 1993). The enhanced call themselves "Humans 2.0"; they claim the unenhanced are but an inferior version. They beg you to enhance. What shall you do? What shall you make of your epistemic predicament? You cannot grasp the contents of the superintelligent beings' thoughts without significant upgrades. But what if their way of thinking is flawed to begin with? In that case, upgrading will surely not help. Is there some sort of neutral vantage point or at least a set of plausible principles with which to guide you in framing a response to such a challenge?

This scenario is merely one example of the kind of issues that will come to the fore as machines outsmart humans, and as some humans themselves enhance their intelligence in ways that allow them to outthink ordinary humans, at least in certain domains. Understanding how to approach such situations requires fruitful collaboration between philosophers of mind, epistemologists, AI specialists, and others.

2.2.1 The ethics of brain enhancement decisions

Should we embrace postbiological intelligence? Enhancement decisions will require deep deliberation about metaphysical and ethical questions that are both controversial and difficult to solve: questions that require reflection about personal identity and the nature of mind, among other issues, and which draw from empirical work in cognitive science. As we explain below, enhancing by moving from carbon to silicon may not be something that preserves your conscious experience or personal identity. Given this, a precautionary stance suggests that we should not enhance unless it is confirmed that consciousness is preserved. For the enhancement is supposed to increase the quality of your life, enabling your survival and giving you more time on the planet as a subject of experience. However, in contrast to a precautionary stance is an attitude of "metaphysical daring" (Mandik 2015). Being metaphysically daring involves making a kind of bet about metaphysical issues such as whether a naturally originating mind could have its consciousness or identity preserved across a transformation from tissue to silicon chips. Metaphysically daring future humans and posthumans may reap the benefits of an enhanced substrate. Indeed, as Mandik argues, systems that exhibit high degrees of metaphysical daring may, in making many more copies of themselves in the form of digital backups, be more fit in a Darwinian sense than

their more cautious evolutionary competitors. Of course, part of what makes the attitude *daring* is the lack of certainty about whether it is correct that such benefits are forthcoming (Mandik 2015).

Given both the lack of certainty on such matters and their life-or-death nature, a pluralistic society should recognize the diversity of different philosophical views on these matters, including a wide range from the metaphysically daring to the metaphysically cautious, and not assume that science itself can answer questions about whether radical forms of brain enhancement are justifiable, or are even compatible with survival, given different views on personal identity and the nature of mind. A good place to further illustrate these observations is with a very extreme "enhancement" case that has been in the news a good deal recently: mind uploading.

3. Mind uploading ("Whole Brain Emulation")

Science fiction has long depicted scenarios in which a person in distress, such as Johnny Depp's character in *Transcendence*, uploads his or her brain in last ditch effort to avoid death. The idea behind uploading is that the person's brain is scanned, and a software model of it is constructed that is so precise that, when run on ultra-efficient hardware, it thinks and behaves in exactly the same way as the original brain. The process of scanning will likely destroy the original brain, as in *Transcendence*, although non-destructive uploading has also been discussed as a more distant possibility (Blackford and Broderick 2014). Uploading is akin to migration to the cloud, but it can be more rapid fire, bypassing your cyborgization. Uploaded beings can be computationally identical to the original human, but they could also become vastly smarter, and less like an ordinary human, as with *Transcendence*.

You might think that if uploading could be developed, day-to-day life would be drastically improved. For instance, on Monday at 6:00 PM, you could have sushi in Tokyo; by 7:30 PM, you could be sipping wine nestled in the hills of the Napa Valley; you need only rent a suitable android body in each locale. Airports could become a thing of the past. Bodily harm matters little to you, for you just pay a fee to the rental company when your android surrogate is injured or destroyed. Formerly averse to risk, you find yourself skydiving and climbing Everest. You think: if I continue to backup, I will live forever. What a surprising route to immortality.

Oxford University's *Future of Humanity Institute* has a brain emulation project that is taking the first steps toward developing uploading. The *OpenWorm* project has successfully uploaded a worm (*C elegans*) and downloaded it to a Lego robot, which behaved like a worm. Uploading could be perfected during a technological singularity. So suppose, like Will Caster, Johnny Depp's character in *Transcendence*, you have just learned you have only a few weeks to live. You recall Steven Hawking's remark: "I think the brain is like a programme . . . so it's theoretically possible to copy the brain onto a computer and so provide a form of life after

death" (Collins 2013). So you wonder: could I truly transfer my consciousness to a computer?

Metaphysics has now become a matter of life and death for you. Would you survive? Philosophers, such as Nick Bostrom and David Chalmers, tend to respond with guarded optimism. But let's consider a literary example to see if even guarded optimism is well-founded. In Robert Sawyer's novel *Mindscan* the protagonist, Jake Sullivan, tries to upload to avoid dying of a brain tumor. He undergoes a non-destructive uploading procedure, and although the contents of his brain are copied precisely, he wakes up after the procedure, still on the operating table, and is astonished to find that he is still stuck in his original body. His consciousness did not "transfer"! Sullivan should have read the personal identity literature in metaphysics, which asks: in virtue of what do you survive over the time? Having a soul? Being a material being? Having the same memories and thought patterns as your earlier self? In deciding whether you could survive uploading, it is important to consider the metaphysical credentials behind each of these views (Schneider 2009). (See also Chapter 5 on personal identity.) AuQ9

One reason Jake should have been suspicious is that objects generally follow a continuous trajectory through space over time – but here, for Jake to "transfer" to his upload, his brain would not even move, and his consciousness would somehow travel inside a computer and then, at a later point, be downloaded into an android. And the stuff that makes up the new Jake would be entirely different. Further, an upload can be downloaded to multiple places at once. But, plausibly, at most only one of these creatures would really be Jake. Which one? Finally, notice that Jake survived the scan. So why believe that any of the uploads is him, rather than the original Jake? In the macroscopic world around us, single objects do not reside in multiple locations at once.

At best, so-called mind uploaders merely create computational copies of themselves that are forms of artificial intelligence (AI). But a copy is not the same as the original. It's a *copy* (Schneider 2014). But if uploads are copies, why be confident, to go back to our original case of your migration to the cloud, that moving to the cloud really preserves your identity? Of course, maybe Derek Parfit is correct. Perhaps there is no identity to begin with (Parfit 1984). In this case, survival is not an issue for you. You may opt to upload for other reasons though – perhaps you believe that creating a psychological duplicate is somehow beneficial.

Or maybe there really is survival, but we are like programs, which can be uploaded and downloaded? In this latter case, maybe uploading can preserve identity because the mind is a program. A program is abstract, like a musical score or equation, and is not a concrete object like a coffee cup, a brain, or a chair. On this sort of view, minds, as programs, are abstract in the sense that the plot of a novel or a song's melody is abstract. If an author emails their latest novel to their publisher, and the publisher prints thousands of copies of the novel, there's only one story here, not thousands. If human minds are abstract in this sense, then the scenario of nond-estructive uploading involves only a single mind, just as there can be a thousand bound copies of a single novel (Mandik 2015, 146–147).

What case can be made for regarding minds as abstract? As Mandik points out:

> Much of what we think, want, and experience is abstract. I can think that there's a dog chasing a cat without there being some particular dog or particular cat that I am thereby thinking about. As Quine (1956) points out, the desire I express in saying "I want a sloop" can just be me wanting relief from slooplessness without there being some certain sloop that I want. Regarding experiences and "what it is like" to have them: I can experience a patch of red on separate occasions, and what it is like to have the experience on the one occasion may be exactly like what it is like on the other occasion. Tye (1995) characterizes all phenomenal character as "abstract" in this sense. If what matters for having my mind is something that can be characterized as abstract in these ways, the possibility opens of a deep analogy between a human life and the story of a novel.
>
> (Mandik 2015, 147)

The view that the mind is abstract in a way that would allow for continuity through uploading is not without its opponents. For instance, Schneider has argued the mind is not a program. For a program or algorithm is like an equation and is abstract. In the fields of philosophy of mathematics and metaphysics, abstract entities are by definition non-spatial, non-causal, atemporal, unchanging, and non-physical. We can tell introspectively that time passes, so minds are temporal, and minds (or more specifically, mental property tokenings, or mental events) are causal, and, relatedly, they experience chance. An equation or algorithm is not located anywhere – although inscriptions and program instantiations are. Our minds and thoughts have concrete locations in space. At best, the mind is a *program instantiation*, which is a concrete entity – a physical object (Schneider, forthcoming).

Regardless of whether we regard the survival conditions of minds as more like the survival conditions for ordinary physical objects or instead like abstract entities such as songs or stories, the important thing is that these are all very controversial positions, relying on certain convictions about the nature of the self, and they militate for different decisions about radical brain enhancement.

As the 21st century unfolds, enhancement decisions will not merely require scientific information about whether uploading can be developed, or whether various minicolumns in your brain can be replaced with silicon implants. They will require philosophical deliberation about the nature of self and mind.

We will revisit radical brain enhancement shortly, for we have yet to explore the important question of whether a silicon being, whether it be you or merely an uploaded copy of you, could be conscious.

4. The hard problem of AI consciousness

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. Chalmers's hard problem of consciousness asks: why does all this information processing in the human brain, under certain conditions, have a felt quality to it? Why aren't we "zombies" in the philosopher's sense, being creatures that lack inner experience (Chalmers 2008)?

As Chalmers emphasizes, this problem 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 (Chalmers 2008). Of course these scientific problem 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" concerning machine consciousness, if you will:

> **The Hard Problem of AI Consciousness:** Would the processing of a silicon-based superintelligent system feel a certain way, from the inside?

A sophisticated AI could solve problems that even the brightest humans are unable to solve, but still, being made of a different substrate, would its information processing feel a certain way from the inside (Chalmers 2008; Searle 1980; Schneider 2015)?

This is not just Chalmers's hard problem applied to the case of AI. 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 AI Consciousness asks *whether* AI, being silicon-based, is even capable of consciousness. It does not presuppose that AI is conscious. These are different problems, but they are both hard problems in their own right – problems that science alone cannot answer.

Ned Block has raised a similar problem, which he calls "The Harder Problem of Consciousness" (Block 2002; McLaughlin 2003). In essence, Block focuses on the case of a "superficial functional isomorph" (SFI) of a human – a being "that is functionally isomorphic to us with respect to those causal relations among mental states, inputs, and outputs that are specified by 'folk psychology'" (Block 2002, 399). According to Block, a SFI need not be conscious, because for all we know, the capacity for consciousness may depend upon a system's underlying substrate, and a silicon-based functional isomorph may lack the right substrate (Block 2002). Block aptly calls attention to the epistemic difficulty of determining whether a different realization would be conscious.

Is our problem just Block's "Harder Problem of Consciousness" then? Block develops his line of thought by focusing on a case of a SFI. In contrast, our hard problem of AI consciousness applies to systems that are not reasonably considered functional duplicates of us, by either armchair folk psychological attributions or scientific functionalist assessments (i.e., psychofunctionalism). It applies to systems that are incredibly different from us with respect to their cognitive and perceptual capacities, such as superintelligences or AGIs not designed to be humanlike. Further, Block's problem arises only for proponents of what he calls "Phenomenal Realism," a view that counts among its commitments that no "a priori or at least armchair analyses of consciousness (or at least armchair sufficient conditions) are given in non-phenomenal terms, most prominently in terms of representation, thought or function." In contrast, our problem can be raised while being neutral about the ultimate status of such analyses. For all we know, there is some as yet unforeseen but correct armchair analysis of consciousness in terms of information processing functions. We are nonetheless currently in the position to be deeply perplexed about *whether* an AI performing such functions would thereby be conscious.

The problem is more general than Block's problem then: simply put, silicon may not be the right medium for consciousness.

Our problem is also related to biological naturalism, a position that is commonly associated with John Searle that has historically denied that AI can be conscious (see Searle 1980). But unlike Searle, we do not find the Chinese Room thought experiment compelling (see Schneider 2015 and Mandik 2017 for discussion).[4] We do not wish to *deny* that machines can be conscious. Instead, we consider it an *open question* whether silicon-based beings can be conscious.

We gain a better understanding of the hard problem of AI consciousness by asking: what considerations may be fueling this problem? Perhaps the problem is fueled, at least in part, by a kind of other minds problem, applied to the case of machines. The case of machines is certainly more challenging, because in the human case, we feel others are minded because of their behavior as well as the fact that they have a physiology that is similar to ours. The case with machines is more challenging, because of a lack of physiological similarity, and it gets quite difficult if a machine's cognitive and perceptual systems are not even loosely similar to our own, as we may not even have similar behaviors to go on.

An other-minds problem, on its own, may fuel the problem, but it does not strike us as being a compelling reason to deny consciousness to AGIs or SAIs. Ethical considerations suggest that it is best to be charitable in these cases, for any mistake could wrongly influence the debate over whether such creatures might be worthy of special ethical consideration as sentient beings. As Asimov's robot stories illustrated, any failure to be charitable to AI could come back to haunt us, as they may treat us as we treated them. Indeed, AIs could pose a "hard problem of carbon-based consciousness" about us, asking if biological, carbon-based entities have the right substrate for experience. After all, how could AI ever be certain that we are conscious?

The Problem of Other Minds is not the only concern that fuels the Hard Problem of AI Consciousness, however.[5] A further, related concern is the following. 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 the field of 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 whether silicon gives rise to consciousness, even if they do not hinder silicon's ability to process information in a superior manner. This is not an endorsement of biological naturalism, but is a consideration indicating that it is not yet clear whether AI can be conscious.

If the answer to the AI hard problem is that silicon cannot be the basis for consciousness, then superintelligent machines – machines that may even one day 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 Caleb that she was in love with him, so too, a clever AI may convincingly behave as if it is conscious.

Further, if subsequent reflection on the AI hard problem reveals that even beings with artificial brains that are computationally like those of humans cannot be conscious, then, in an extreme, horrifying case, humans upload, 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. Even the slightest chance that this could happen should give us reason to proceed in the development of uploading and brain implant technologies with caution. These issues urgently need to be addressed.

4.1 A solution?

Is there a means to answer the AI Hard Problem? Two scenarios are suggestive.

First, although it is unlikely, we could find silicon-based *natural* intelligence on a planet – silicon-based life that arose through chemical processes, rather than being constructed by a biological species. If these creatures have a phenomenological vocabulary – a vocabulary of what it is like to experience the world – it would not be due to their being programmed by a biological species to act as if they had experience. Further, their phenomenological vocabulary cannot be a mere mimicry of the behavior or vocabulary of a biological species that evolved separately and had contact with them. What we need is pure, untainted silicon phenomenology, if you will.

If untainted naturally occurring silicon-based phenomenology was discovered, this would make more plausible the claim that artificial silicon-based systems could support phenomenology. Of course, even in this case some may still doubt whether artificial systems could be conscious (based, for instance, on

considerations about teleofunction or John Searle's alleged derived/non-derived intentionality distinction) (cite).

AuQ10

Let's turn now to a second suggestion for making progress on the Hard Problem of AI Consciousness. Let us return to the case of one's migration to the cloud. In the process of migrating, neurons that form the neural basis of one's consciousness are gradually replaced by silicon chips. If, during this process, a prosthetic part of the brain ceases to function normally – specifically, if it ceases to give rise to the aspect of consciousness that that brain area is responsible for – then there should be behavioral indications, including verbal reports. An otherwise normal person should be able to detect, or at least indicate to others through odd behaviors, that something is amiss, as with traumatic brain injuries involving the loss of consciousness in some domain, such as blindsight or blindness denial. This would indicate a "substitution failure" of the artificial part for the original component.

But should we really draw the conclusion, from a substitution failure, that the underlying cause is that silicon cannot be a neural correlate of conscious experience? Why not instead conclude that scientists failed to program in a key feature of the original component – a problem which science can eventually solve? But after years and years of trying, we may reasonably question whether silicon is a suitable substitute for carbon when it comes to consciousness. This would be a sign that the answer to the hard problem of AI consciousness is negative: AI cannot be conscious. But even a longstanding substitution failure would not be *definitive*, for there is always the chance that our science has fallen short. But this scenario would provide some evidence for a negative answer.

Readers familiar with Chalmers's "absent qualia, dancing qualia" thought experiment may object that we've missed something, for Chalmers's thought experiment supports the view that consciousness supervenes on functional configuration: if you fix the psychofunctional facts, you fix the qualia. But we are disputing that functional isomorphism occurs in the first place. We consider it an open question.

If silicon systems cannot be conscious, then the functional facts cannot be fixed. When it comes to consciousness, carbon and silicon are not functionally interchangeable. For why would a silicon system, S2, be a psychofunctional isomorph of the original system, S1, after the transfer? S2s replaced brain region, or minicolumn, being made of silicon, will always differ causally from the replaced component. For wouldn't the new silicon component somehow signal to other brain areas that there is a defect in consciousness, as with neurophysiological deficits?

Could the silicon chip be doctored, so as to signal consciousness when consciousness was absent though? This is a tricky question. It could be the case that there are some observational false positives, in which case, we may fail to rule out certain cases of non-conscious systems. But would it then be a genuine functional isomorph of a carbon system? It is not clear that it would be, for the brain chip would need to prevent signaling to other brain areas that consciousness is lacking.

The conscious system would not. Our example does not require rejecting the view that qualia supervenes on functional organization, then.

Conclusion

The practical and intellectual challenges we foresee philosophers of mind helping to meet have here fallen into four groups. The first group of challenges centered on the possibility of superintelligent artificial intelligence, a technology that may potentially populate our world with nonhuman selves bestowed with capacities that meet or exceed our own. The second group of challenges concern brain enhancement, extreme cases of which might result in beings more posthuman than human. Even more extreme transformations formed the core of the third group of challenges, those that centered on the hypothetical technology of mind uploading, which might constitute a way for human minds to survive indefinitely through digital backup, or might instead be merely a very expensive form of suicide. Fourth and finally, we raised the hard problem of AI consciousness, a special form of the problem of determining whether a given entity is such that there's something it feels like to itself "from the inside." There's an ethical element to this problem, for we recognize an ethical imperative not to inflict avoidable suffering upon any being, whether they be natural or artificial.

We surely have just scratched the surface in exploring ways that philosophy of mind can help shape the future. Despite the numerous ways that will surely escape our foresight, we are confident that the technological changes that await us, in particular those involving information processing technology, will pose problems that science alone cannot equip society to solve.

Notes

1 Further, there is growing concern among policymakers and the public that AI will eventually outmode humans, leading to technological unemployment (Frey and Osborne, 2013).
2 *DeepMind* is a British artificial intelligence company acquired by Google in 2014. The IBM's *Watson* program is a natural-language processing computer system that famously competed on *Jeopardy!* in 2011.
3 Worries about technological unemployment do not assume that AGIs will be superintelligent; indeed, people can become unemployed due to the development of domain-specific AI systems that are not even at the level of being AGIs.
4 In Searle's famous Chinese Room argument, Searle appeals to the thought experiment of the "Chinese Room" to argue against the possibility of artificial systems being genuinely intelligent. In the thought experiment, Searle runs a program for understanding Chinese despite himself understanding only English. Observers outside of the Chinese room send and receive messages to and from the room that lead them to believe the room's inhabitant is perfectly conversant in Chinese. But Searle is orchestrating the message exchange solely in virtue of following instructions written in English.
5 For discussion of the Chinese Room Thought experiment, see (Schneider 2015 and Mandik 2017).

Bibliography

Bennett, J., and Shostak, S. (2012). *Life in the Universe*, 3rd ed., Boston: Addison-Wesley.

Blackford, R., and Broderick, D., (2014). *Intelligence Unbound: The Future of Uploaded and Machine Minds*. Boston: Wiley Blackwell.

Block, N. (2002). "The Harder Problem of Consciousness," *The Journal of Philosophy*, XCIX: 1–35.

Bostrom, D. (2017). The Transhumanist FAQ: v 2.1. World Transhumanist Association.

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

Chalmers, D. J. (2010). "The Singularity: A Philosophical Analysis," *Journal of Consciousness Studies*, 17, 7–65.

Chalmers, D. J. (2008). "The Hard Problem of Consciousness." in Velmans, Max and Schneider, Susan (eds.), *The Blackwell Companion to Consciousness*. Oxford: Wiley-Blackwell.

Collins, N. (2013). "Hawking: 'in the Future Brains Could Be Separated From the Body,'" *The Telegraph*, 20 September.

Corabi, J., and Schneider, S. (2012). "The Metaphysics of Uploading," *Journal of Consciousness Studies*, 19.

Frey, C. and and Osborne, M. (2013). https://www.oxfordmartin.ox.ac.uk/downloads/academic/future-of-employment.pdf

Garreau, J. (2004). *Radical Evolution: The Promise and Peril of Enhancing our Minds, our Bodies – and What it Means to be Human*. New York: Doubleday.

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

Mandik, P. (2015). Metaphysical Daring as a Posthuman Survival Strategy, *Midwest Studies in Philosophy*, 39(1), 144–157.

Mandik, P. (2017). "Robot Pain," in Corns, J. (ed.), *The Routledge Handbook of Philosophy of Pain*. New York: Routledge: 200–209.

McGinn, C. (1993). *Problems in Philosophy*. Oxford: Oxford University Press.

McLaughlin Brian (2003). P. A Naturalist-Phenomenal Realist Response to Block's Harder Problem, *Philosophical Issues* 13, 163–204.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63 (2), 81–97. *doi:* 10.1037/h0043158. PMID 13310704.

Müller, V. C., and Bostrom, N. (2014). "Future Progress in Artifi-cial Intelligence: A Survey of Expert Opinion," in Müller, Vincent C. (ed.), *Fundamental Issues of Artificial Intelligence*. Synthese Library. Berlin: Springer.

Parfit, D. (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Quine, W. V. O. (1956). Quantifiers and Propositional Attitudes, *Journal of Philosophy*, 53, 177–187.

Schneider, S. (2009a). "Mindscan: Transcending and Enhancing the Human Brain," in Schneider, Susan (ed.), *Science Fiction and Philosophy*. Oxford: Blackwell Publishing 241–255.

Schneider, S. (2009b). "Science Fiction Thought Experiments as a Window into Philosophical Puzzles," in *Science Fiction and Philosophy*. Chichester: Wiley-Blackwell.

Schneider, S. (2011). *The Language of Thought: A New Philosophical Direction*. Boston: MIT Press.

Schneider, S. (2014). "The Philosophy of 'Her'," *The New York Times*, 2 March.
Schneider, S. (2015). "Alien Minds," *Discovery* (an astrophysics trade anthology, based on a NASA/Library of Congress Symposium), Steven Dick, Cambridge: Cambridge University Press.
Schneider, S. (forthcoming). "The Mind is not the Software of the Brain (Even if it is Computational)", ms.
Searle, J. (1980). Minds, Brains and Programs, *The Behavioral and Brain Sciences*, 3, 417–457.
Tye, Michael (1995). *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press.
Wallach, W. and Allen, C. (2010). *Moral Machines*. Oxford: Oxford University Press.