**The Future of the Mind**

Susan Schneider

**Contents**

Introduction: Designing the Mind

Chapter 1: The Age of AI

Chapter 2: The Problem of AI Consciousness

Chapter 3: Consciousness Engineering

Chapter 4: How to Catch an AI Zombie: Testing for Consciousness in Machines

Chapter 5: Could you Merge with AI?

Chapter 6: Getting a Mindscan

Chapter 7: A Universe of Singularities

Chapter 8: Is the Mind a Software Program?

Chapter 9: Conclusion: The Afterlife of the Brain

Appendix One: On Transhumanism

Note to readers: Thank you very much for your interest in my manuscript. Please note that some citations are unfinished and the footnotes editing. (A copyeditor will convert them to a uniform format.) While this ms. circulates, I am adding citations to the text. I’m sorry parts that are rough around the edges.

**Introduction**

***It is 2045. Today, you are shopping. Your first stop is The Center for Mind Design. As you walk in, a large menu stands before you. It lists brain enhancements with funky names, like “Human Calculator,” (a brain chip that gives you savant-like mathematical abilities), “Hive Mind” (a chip allowing you to experience the innermost thoughts of your loved ones), “Zen Garden” (a microchip for Zen master-level meditative states) and so on. What would you select, if anything? Enhanced attention? Mozart-level musical abilities? You can order one, or a bundle of several.***

***Later, you visit the android shop. It is time for that new android to take care of the house. The menu of AI minds is vast and varied. Some AIs have databases that span the internet and heightened perceptual abilities, but like those with savant syndrome, they lack the intelligence of a normal human. Others have senses humans lack altogether. You carefully select the options that suit your family. Today is a day of mind design decisions.***

**This book concerns the future of the mind. It’s about how our understanding of ourselves, our minds, and our nature can drastically change the future, for better or for worse. Discussions of the future of AI are all too often mired in philosophical misunderstandings about what AI can and cannot achieve. If we humans ever do make mind design decisions, we must proceed carefully.**

**Our brains evolved for specific environments and are greatly constrained by anatomy and evolution. But artificial intelligence has opened up a vast design space, offering new materials and modes of operation, as well as new ways to explore the space at a rate much faster than biological evolution. I call this exciting new enterprise *mind design*. Mind design is a form intelligent design, but we humans, not God, are the designers.**

**I find the prospect of mind design humbling because frankly, we are not all that evolved. As the alien in the Carl Sagan film, *Contact,* says, upon first meeting a human, “You're an interesting species. An interesting mix. You're capable of such beautiful dreams, and such horrible nightmares.”**[[1]](#endnote-1) **We walk the moon, we harness the energy of the atom, yet violence, hunger and racism are not uncommon. The contrast is heart wrenching.**

**It might seem less worrisome, when, in contrast, I tell you, as a philosopher, that we are utterly confounded about the nature of the mind. But there is a cost to not understanding issues in philosophy too, as you’ll see when you consider the two central threads of this book.**

**The first central thread is something quite familiar to you. It has been there throughout your life: your consciousness. Notice that as you read this, it feels like something to be *you*. You are having bodily sensations, you are seeing the words on the page, and so on. Consciousness is this felt quality to your experience. Without consciousness, there would be no pain or suffering, no feeling of joy, no burning drive of curiosity, no pangs of grief. Experiences, positive or negative, simply wouldn’t exist.**

**It is as a conscious being that you long for vacations, hikes in the woods, or spectacular meals. Because consciousness is so immediate, so familiar, it is natural that you primarily understand consciousness through your own case. After all, you do not need to read a neuroscience textbook to understand what it feels like, from the inside, to be conscious. Consciousness is essentially this kind of inner feel. It is this kernel –your conscious experience – which, I submit, is characteristic of having a mind.**

**Now for some bad news. The second central thread of the book is that failing to think through the philosophical implications of AI could lead to the failure of conscious beings to flourish.**  If we are not careful we may experience one or more *perverse realizations* of AI technology, that is, situations in which AI fails to make life easier but instead leads to our own suffering or demise, or to the exploitation of other conscious beings.

Many have discussed AI-based threats to human flourishing. These threats range from hackers attacking the power grid to superintelligent autonomous weapons that seem right out of the movie, *The Terminator*. In contrast, the issues I raise have received less attention. Yet they are no less significant. These perverse realizations I have in mind generally fall into two categories: (i), situations involving the creation of conscious machines and (ii), scenarios that concern radical brain enhancement, such as the enhancements at the hypothetical *Center for Mind Design*. Let’s consider each of these issues in turn.

***Conscious Machines?***

**Suppose that we create sophisticated, general purpose AIs: AIs that flexibly connect issues from different domains, and which even rival humans in their capacity to reason. Would we, in essence, be creating mentality in machines -- conscious machines that are both selves and subjects of experience?**

**When it comes to machine consciousness, we are in the dark.** One thing is clear, however: The question of whether AIs could have experience is key to how we value their existence. Consciousness is the philosophical cornerstone of our moral systems, being key to our judgment of whether someone or something is a self or person rather than a mere automaton. After all, would you really be comfortable giving that android shop your business, if the items on the menu were conscious beings?

If I were an AI director at *Google* or *Facebook*, thinking of future projects, I wouldn’t want the ethical muddle of inadvertently designing a conscious system. Developing a system that turns out to be conscious or sentient could lead to accusations of AI slavery and other public-relations nightmares. It could even lead to a ban on the use of AI technology in certain sectors.

I’ll suggest that **all this may lead AI companies to engage in *consciousness engineering*—a deliberate engineering effort to avoid building conscious AI for certain purposes, while designing conscious AIs for other situations, if appropriate. Of course, this assumes consciousness is the sort of thing that can be designed in and out of systems. Consciousness may be an inevitable byproduct of building an intelligent system, or it may be altogether impossible.**

Some suspect that AI will be the next phase in the evolution of intelligence on Earth. You and I, how we live and experience the world right now, are just an intermediate step, a rung on the evolutionary ladder. For instance, Stephen Hawking, Nick Bostrom, Elon Musk, Bill Gates and many others have raised “the control problem,” a problem of how humans can control their own AI creations, if the AIs outsmart us. Suppose we create AIs that have human-level intelligence. It has self-improvement algorithms, and with rapid computations, it quickly discovers ways to become vastly smarter than us, becoming a superintelligence, that is, an AI that outthinks us in every domain. Because it is superintelligent, we probably can’t control it. It could, in principle, render us extinct.[[2]](#endnote-2) This is only one way that synthetic beings could supplant organic intelligences—humans may slowly merge with AI through cumulatively significant brain enhancements.

It is important to put issue into a larger, universe-wide context. In my NASA funded project, I urged that a similar phenomenon could be going on other planets as well; elsewhere in the universe, other species may be rungs on their own evolutionary ladders. As we search for life elsewhere, we must bear in mind that the greatest alien intelligences may be *postbiological*, being AIs that evolved from biological civilizations.

But if it doesn’t feel like anything to be an AI, we have to ask whether we want to be a mere rung on an evolutionary ladder. Is a universe full of intelligent, non-conscious machines really more valuable than conscious life, like us?  **In this case, even if the universe is stocked full of superintelligent AIs, no conscious beings would remain to enjoy the world.**

Consciousness may also be key to how AI values *us*. The value an AI places on us may very well hinge on whether it has an inner life; using its own subjective experience as a springboard, it could recognize in us the capacity for conscious experience. After all, to the extent we value the lives of nonhuman animals, we tend to value them because we feel an affinity of consciousness—thus most of us recoil from killing a chimp, but not from eating an orange. Conscious AI could help with the control problem, leading to safer AI.

When it comes to the issue of sentient or conscious machines, this is just the tip of the iceberg, as you’ll see. For instance, I will also explore ways to determine if synthetic consciousness exists, outlining tests I’ve developed at the Institute for Advanced Study in Princeton. If AI consciousness is as significant as I claim, we’d better know if we’ve built it.

Now let’s consider the suggestion that humans should merge with AI. Suppose that you are at *The Center for Mind Design*. What would you order from the menu, if anything? Having read these first few pages, you are probably already getting a sense that mind design decisions are not so simple.

*Could You Merge with AI?*

**I wouldn’t be surprised if you find the idea of augmenting your brain with microchips wholly unnerving. If companies like Facebook and Cambridge Analytica cannot even respect our privacy right now, think of the potential for abuse if your innermost thoughts are encoded on microchips, perhaps even being accessible somewhere on the Internet.**

**But let’s play suppose AI regulations improve, and our brains could be protected from hackers and corporate greed. Perhaps you begin to feel the pull of enhancement, as others around you appear to benefit from the technology. After all, if merging with AI leads to superintelligence and radical longevity, isn’t it better than the alternative—i.e., the inevitable degeneration of the brain and body?**

**The idea that humans should merge with AI is very much in the air these days, being offered both as a means for humans to avoid being outmoded by AI in the workforce, and as a path to superintelligence and immortality. For instance, Elon Musk recently commented that humans can escape being outmoded by AI by “having some sort of merger of biological intelligence and machine intelligence.” To this end, he’s founded a new company, *Neuralink*. One of its first aims is to develop “neural lace”, an injectable mesh that connects the brain directly to computers. Neural lace and other AI-based enhancements are supposed to allow data from your brain to travel wirelessly to one’s digital devices or to the cloud, where unlimited computing power is available.**

**Musk’s motivations may not be less than altruistic though. He is pushing a product line of AI enhancements, products that presumably solve a problem that the field of AI itself created. Perhaps these enhancements will turn out to be beneficial, but to see if this is the case, we will need to move beyond all the hype. Policymakers, the public, and even the AI researchers themselves need a better idea of what is at stake.**

**For instance, if AI cannot be conscious, if microchips replace parts of the brain responsible for consciousness, your “enhancement” would end your life as a conscious being. You’d become what philosophers call a “zombie” -- a non-conscious simulacra of your earlier self.**

**And even ifmicrochips can replace parts of the brain responsible for consciousness, radical enhancement is still a major risk. Given what we know about the nature of the self, after too many changes, the person who remains may not be you. Each human who enhances, unbeknownst to them, may end their life in the process.**

**In my experience, many proponents of radical enhancement fail to appreciate that the enhanced being may not be you. For they tend to sympathize with a conception of the mind that says the mind is a software program. Just as you can upload and download a computer file, so too, your mind, as a program, could be uploaded onto the cloud. You can enhance in radical ways, and still be the same “program.” This is a technophile’s route to immortality – the brain’s new “afterlife,” if you will. As alluring as a technological form of immortality may be, we’ll see that this view of the mind is deeply flawed.**

So, if, decades from now, you stroll into a mind design center, or visit an android store, remember, the **AI technology you purchase could fail to do its job for philosophical reasons.**  *Buyer beware*. But before we delve further into this, you may suspect that these issues will forever remain hypothetical. For I am wrongly assuming that sophisticated AI will developed. Why suspect any of this will happen?

**Chapter One: The Age of AI**

You may not think about AI on a daily basis, but it is all around you. It’s here when you do a Google search. It’s here beating the world *Jeopardy!* and *Go* champions. And it’s getting better by the minute. But we don’t have general purpose AI yet -- AI that is capable of holding an intelligent conversation on its own, thinking flexibly about various topics, and even, perhaps, outthinking us. This sort of AI is depicted in films like *Her* and *Ex Machina*, and it may strike you as the stuff of science fiction.

I suspect it’s not that far away, though. The development of AI is driven by market forces and the defense industry – billions of dollars are now pouring into the development of smart household assistants, robot super-soldiers, and supercomputers that mimic the workings of the human brain. Indeed, the Japanese government has launched an initiative to have androids take care of the nation’s elderly, in anticipation of a labor shortage.



[Pictured above is an android designed by Hiroshi Ishugaru, which sits next to the human it was modeled after—Ishugaru’s daughter.]

**Given the current rapid-fire pace of its development, AI may soon advance to artificial general intelligence (“AGI”) within the next decades. AGI is intelligence that, like human intelligence, can combine insights from different topic areas and display flexibility and common sense. Indeed, AI is already projected to outmode many human professions within the next several decades.** According to a survey, for instance, the most cited AI researchers expect AI to “carry out most human professions at least as well as a typical human” within a 50% probability by 2050, and within a 90% probability by 2070 (Müller & Bostrom, 2016).

I’ve mentioned that many have warned of the rise of “superintelligent” AI: synthetic intelligences that outthinks the smartest humans in every domain, including common sense reasoning and social skills. Superintelligence could destroy us, they urge. In contrast, Ray Kurzweil, a futurist who is now a director of engineering at *Google*, depicts a technological utopia bringing about the end of disease, poverty and resource scarcity. Kurzweil has even discussed the potential advantages of forming friendships with personalized AI systems, like the Samantha program in the film, *Her.*

*The Singularity*

Kurzweil and other transhumanists contend that we are fast approaching a “technological singularity,” a point at which AI far surpasses unenhanced human intelligence, and is capable of solving problems we weren’t able to solve before, with unpredictable consequences for civilization and human nature.

The idea of a singularity is from work on the physics of black holes. Black holes are described as “singular” objects in space and time; places where normal physical laws break down. All bets are off. The singularity is projected to cause runaway technological growth and massive changes to civilization. Beyond this “event horizon” technological innovation, the human mind, and indeed, all of human history, will change immensely. Superintelligent AI -- AI that outthinks us in every dimension – is the driving force behind the technological singularity.

**I am not sure that humans will lose control of AI during a singularity, as some say. This depends on whether humans themselves enhance with AI technologies, so they can keep abreast of it. And it may be that the technological innovations are not so rapid-fire that they lead to a full-fledged singularity, in which the world changes almost overnight. But this shouldn’t distract us from the larger point:** we must come to grips with the possibility that as we move further into the 21st Century, unenhanced humans may not be the most intelligent beings on the planet for that much longer. The greatest intelligences on the planet may be synthetic.

**Indeed, I think we already see reasons why synthetic intelligence will outperform us. Even now, silicon microchips seem to be a better medium for information processing than groups of neurons. Neurons reach a peak speed of about 200 hertz. As I write this chapter, the worlds’ fastest computer is the Summit supercomputer at the Oak Ridge Laboratory in Tennessee.** Summit’s speed is *two hundred petaflops*-that’s two hundred million billion calculations per second. To match this speed, all the people on Earth would need to do a calculation every moment of every day for three hundred and five days. This is what Summit can do in the blink of an eye.[[3]](#endnote-3)

**Of course, speed is not everything. Your brain is the product of 3.8 billion years of evolution (the estimated age of life on the planet). Overall, your cognitive and perceptual performance exceeds Summit’s. your cleverness isn’t merely a matter of computational speed, it is a matter of the way your cognitive system is organized. But AI has almost unlimited room for improvement. It may not be long before a supercomputer can be engineered to match or even exceed the intelligence of the human brain through reverse-engineering the brain and improving upon its algorithms, or through some combination of reverse engineering and judicious algorithms that aren’t based on the brain’s workings at all.**

**In addition, an AI can be downloaded to multiple locations at once, is easily backed up and modified, and can survive under conditions that biological life struggles with, including interstellar travel. Our measly brains are limited by cranial volume and metabolism; AI, in stark contrast, could extend its reach across the Internet and even set up a Galaxy-wide “computronium” -- a massive supercomputer that utilizes all the matter within our galaxy for its computations. There is simply no contest. AI could be far more durable than us.**[[4]](#endnote-4)

***The Jetsons Fallacy***

**Remember, AI will not just make for better robots and supercomputers.** Consider *Star War*s and *Westworld*, or consider the classic television cartoon, *The Jetsons*. In these storylines, humans are surrounded by sophisticated AIs, while themselves remaining unenhanced. The historian Michael Bess has called this *The Jetsons Fallacy*. In reality, AI will not just transform the world, it will transform us. Neural lace, the artificial hippocampus, brain chips to treat mood disorders -- these are just some of the mind altering technologies already under development. So, *The Center for Mind Design* is not that far-fetched.

**Increasingly, the human brain is being regarded as something that can be hacked, like a computer. There are already hundreds of projects in the US alone developing brain implant technologies to treat mental illness, motion-based impairments, strokes, dementia, autism, and more. The medical treatments of today will give rise to the enhancements of tomorrow. After all, people long to be smarter, more efficient, or simply have a heightened capacity to enjoy the world. To this end, AI companies like *Google, Neuralink* and *Kernel* are developing ways to merge humans with machines. Within the next several decades, you may be a cyborgs.**

*Transhumanism*

The research is new, but it is worth emphasizing that the basic ideas have been around far longer, in the form of a philosophical and cultural movement known as *transhumanism*. Julian Huxley coined the term “transhumanism” in 1957, when he wrote that in the near future “the human species will be on the threshold of a new kind of existence, as different from ours as ours is from that of Peking man” (Huxley, 1957, pp. 13–17). Max More, a key founder of the transhumanist movement, introduced the term in 1990 in its modern sense. [[5]](#endnote-5)

Transhumanism holds that the human species is now in a comparatively early phase and that its very evolution will be altered by developing technologies. Future humans will be quite unlike their present-day incarnation in both physical and mental respects and will in fact resemble certain persons depicted in science fiction stories. Transhumanists share the belief that an outcome in which humans have radically advanced intelligence, near immortality, deep friendships with AI creatures, and elective body characteristics is a very desirable end, both for one’s own personal development and for the development of our species as a whole.

**Despite its science fiction-like flavor, many of the technological developments that transhumanism depicts seem quite possible: indeed, the beginning stages of this radical alteration may well lie in certain technological developments that either are already here (if not generally available), or are accepted by many in the relevant scientific fields as being on their way (Roco & Bainbridge 2002; J. Garreau 2005). In the face of these technological developments, transhumanists offer an agenda of public import, with ideas being developed at places like Oxford University’s Future of Humanity Institute and at the Institute for Ethics and Emerging Technologies. (To further acquaint the reader with transhumanism, I’ve included the Transhumanist Declaration in the Appendix.)**

**You may be surprised to learn that I consider myself a transhumanist, but I do. I first learned of transhumanism when an undergraduate at the University of California at Berkeley, when I joined the Extropians, an early transhumanist group. After pouring through my boyfriend’s science fiction collection, and reading the Extopian listserve, I was enthralled by the transhumanist vision of a technotopia on Earth. It is still my hope that emerging technologies will provide us with radical life extension, help end resource scarcity and disease, and even enhance our mental lives, should we wish to enhance. The challenge is how to get there from here.**

*****A Few Words of Warning*****

**Whereas transhumanism once seemed like science fiction, a good deal of research is already underway to develop the very technologies transhumanists have been discussing for decades. For instance, Oxford University’s Future of Humanity Institute released a report on the technological requirements for uploading a mind to a machine. A Defense Department agency has funded a program, *Synapse*, that is trying to develop a computer that resembles the brain in form and function. The futurist Ray Kurzweil, now a director of engineering at Google, has even discussed the potential advantages of forming friendships, “Her”-style, with personalized artificial intelligence systems. All around us, science fiction is becoming science fact.**

**Still, it would be absurd to suggest that a book written today could accurately predict the contours of mind design space. I certainly do not aspire to this lofty ambition. Throughout the book, I stress that the underlying philosophical mysteries may not diminish as our scientific knowledge and technological prowess increases. I’m afraid that when it comes to the nature of the self and mind, there are no easy answers.**

**An additional layer of uncertainty is that the future is obviously unpredictable. Indeed, it pays to keep in mind two important ways in which the future is opaque. First, there are known unknowns – the things we know we do not know. For instance, I will suggest that we cannot be sure about the nature of the self and mind, and that this uncertainty should impact the kind of brain enhancement decisions we make. But the most challenging unknowns are the *unknown unknowns*—the things we do not know we do not know. One of my aims is to uncover some known unknowns, but obviously, I can say little about unknown unknowns. They are unknown, by definition.**

**In the next chapters we turn to one of the great known unknowns: the puzzle of conscious experience. We will appreciate how the puzzle arises in the human case, and then, we will ask:** how can we even recognize consciousness in beings that may be vastly intellectually different from us, and which may even be made of different substrates? A good place to begin is by simply appreciating the depth of the issue.

**Chapter Three: The Problem of AI Consciousness**

Consider what it is like to be a conscious being. Every moment of your waking life, and whenever you dream, it feels like something to be you. When you hear your favorite piece of music, or smell the aroma of your morning coffee, you are having conscious experience. While it may seem a stretch to claim that today’s AIs are conscious, as AI grows in sophistication, could these more sophisticated AIs experience the world? Could they have sensory experiences, or feel emotions, like the burning of curiosity or the pangs of grief, or even have experiences that are of an entirely different flavor from our own? Let us call this *The Problem of AI Consciousness.* No matter how impressive AIs of the future turn out to be, if machines cannot be conscious, then they could exhibit superior intelligence, but they would lack inner mental lives.

In the context of biological life, intelligence and consciousness seem to go hand-in-hand. Sophisticated biological intelligences tend to have complex and nuanced mental lives. But would this correlation apply to nonbiological intelligence as well? Many suspect so. For instance, transhumanists, such as Ray Kurzweil, tend to hold that just as human consciousness is richer than that of a mouse, so too, unenhanced human consciousness would pale in comparison to the experiential life of a superintelligent AI.[[6]](#endnote-6) But as we shall see, this line of reasoning is premature. There may be no special androids that have the spark of consciousness in their machine minds, like Delores in *Westworld* or Rachel in *Bladerunner*. And even if AI surpasses us intellectually, we still may stand out in a crucial dimension: it feels like something to be us.

Let’s begin by simply appreciating how perplexing consciousness is, even in the human case.

**AI Consciousness and the Hard Problem**

The philosopher David Chalmers has posed “the hard problem of consciousness,” asking: why does all the information processing in the brain need to feel a certain way, from the inside? As Chalmers emphasized, this problem doesn’t seem to be one that has a purely 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 all the information processing in the visual system. 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 problems are difficult problems in their own right; Chalmers merely calls them “easy problems” to contrast them with the “hard problem” of consciousness, which he thinks will not have a scientific solution.

We now face yet another perplexing issue involving consciousness -- a kind of “hard problem” concerning machine consciousness, if you will:

**The Problem of AI Consciousness**: Would the processing of an AI 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 have a felt quality to it?

The Problem of AI Consciousness is not just Chalmers’ hard problem applied to the case of AI, though. For Chalmers’ hard problem of consciousness assumes that we are conscious. After all, each of us can tell from introspection that we are now conscious. It asks: *why* we are we conscious? Why does some of the brain’s information processing feel a certain way from the inside? In contrast, the problem of AI consciousness asks whetheranAI, being made of a different substrate, like silicon, is even capable of consciousness.[[7]](#endnote-7) It does not presuppose that AI is conscious – this is the question. These are different problems, but perhaps they are both problems that science alone cannot answer.

Discussions of the problem of AI consciousness tend to be dominated by two opposing positions. *Biological naturalists,* on the one hand, claim thateven the most sophisticated forms of AI will be devoid of inner experience (Searle 2008; 1990; Blackmore 2003).[[8]](#endnote-8) The capacity to be conscious is unique to biological organisms, so that even sophisticated androids and superintelligences will not be conscious. A second influential approach, which I’ll simply call “techno-optimism about AI consciousness”, or “techno-optimism” for short, rejects biological naturalism. Drawing from empirical work in cognitive science, it urges that consciousness is computational through and through, so sophisticated computational systems will have experience.

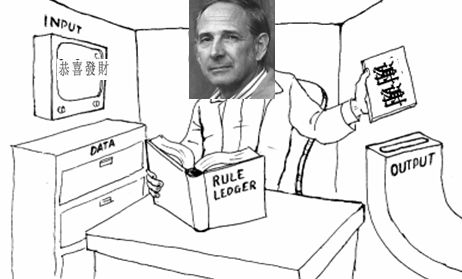
If biological naturalists are correct, then a romance or friendship between a human and an AI, for instance as depicted in the aforementioned film *HER*, would be hopelessly one-sided. The AI may be smarter than humans, and it may even project compassion or romantic interest, much like Samantha, but it wouldn’t have more any experience of the world than your laptop. Moreover, few humans would want to join Samantha in the cloud. To upload your brain to a computer would be to forfeit your consciousness. The technology could be impressive, perhaps your memories could be accurately duplicated on the cloud, but that stream of data would not be you; it wouldn’t have an inner life.

**Biological Naturalism**

Now let us ask: What could motivate biological naturalism? You might suspect that to show that AI cannot be conscious, the biological naturalist should do the following: locate a special property or feature that is responsible for, or at least closely correlated with, biological consciousness, (call it “P”) and which cannot be a property of the particular substrate in which AI is built, due to something about that alternate substrate’s chemical composition. Thus far, P has not been discovered. But what if it is? Is this really what is needed to prove that the AI that we build will not be conscious?

While it would be intriguing to learn that other substrates cannot have P, I would caution against taking this as evidence that a given AI is not conscious. For perhaps it is capable of having a different type of consciousness property, F, which is specific to systems of that substrate. As I shall explain in Chapter Three, to tell whether AI is conscious, we must reach beyond the chemical properties of particular substrates, and seek clues within the AIs behavior.

A different sort of motivation may stem from reflecting on a famous thought experiment, called “The Chinese Room Thought Experiment,” due to the philosopher John Searle. Searle asks you to suppose that he is locked inside a room. Inside the room, there is a small opening through which he is handed cards with strings of Chinese symbols. But Searle doesn’t speak Chinese, although before he goes inside the room, he is handed a book of English rules that allow him to look up a particular string and then write down some other particular string in response. So Searle goes in the room, and he is handed a note card, with Chinese script. He then consults his book, writes down Chinese symbols, and passes the card through a second hole in the wall.



(http://rationallyspeaking.blogspot.com/2013/07/philosophy-not-in-business-of-producing.html)

What does this have to do with AI, you may ask? Notice that from the vantage point of someone outside the room, Searle’s responses are indistinguishable from those of a Chinese speaker. Yet he hasn’t a clue what he’s written means. Like a computer, he’s produced answers to inputs by manipulating formal symbols. The room, Searle, and the cards all form a kind of information processing system, but he doesn’t understand a word of Chinese. So how could the manipulation of data by dumb elements, none of which understand language, ever produce something as glorious as understanding or experience? According to Searle, the thought experiment suggests that no matter how intelligent a computer seems, the computer is not really thinking or understanding. It is only engaging in mindless symbol manipulation.

By the way, it is worth noting that Searle views understanding as being closely related to consciousness, although he doesn’t always make this point explicit. For the sake of argument, let us assume he is right. After all, it isn’t implausible that when we understand, we are conscious; not only are we conscious of the point we are understanding, but importantly, we are in an overall state of wakefulness and awareness.

So, is Searle correct that the Chinese room cannot be conscious? He is, of course, correct that the person who is manipulating symbols in the room doesn’t understand Chinese. But notice that the salient issue, in determining whether the system is conscious, is not whether anyone in the room understands Chinese; the person in the room is merely one part of a larger system that includes cards, a book, a room, and so on. The relevant question is whether the *system as a whole* understands Chinese. The view that the system as a whole truly understands, and is conscious, is a popular response to the Chinese Room. It has become known as the “Systems Reply.”(grab citations from my “Alien Minds” paper)

The Systems Reply strikes me as being right in one sense, while wrong in another. The System’s Reply is correct that the real issue, in considering whether machines are conscious, is whether the system as a whole is conscious, not whether one component is. Suppose you are holding a steaming cup of green tea. No single molecule in the tea is transparent, but the tea is. Transparency is a feature of certain complex systems. **In a similar vein, no single neuron, or area of the brain, realizes, on its own, the complex sort of consciousness that a self or person has. Now, maybe panpsychists are right that there is a miniscule amount of consciousness in every particle in the universe; maybe there is a smidgeon of consciousness (“microconsciousness”) that inheres in each and every particle. (This is an issue I touch upon in chapter X). Even so,** the kind of consciousness that intelligent biological systems possess involves the complex interaction and integration between various parts of the brain, such as the brainstem, claustrum and thalamus. Consciousness emerges from an elaborate dance of elementary particles, a dance which science cannot yet reproduce and doesn’t even understand. It is a feature of highly complex systems, not a homunculus within a larger system akin to Searle standing in the room.

Relatedly,Searle’s reasoning is that the system doesn’t understand Chinese because *he* doesn’t understand Chinese. In other words, the whole cannot be conscious because *a part* isn’t conscious. But this line of reasoning is flawed: we already have an example of a conscious system that understands even though a part of it does not: the human brain. The cerebellum possesses 80% of the brain’s neurons, yet we know that it isn’t required for consciousness, because there are people with unimpaired consciousness who have cerebellar agenesis, being born without a cerebellum. I bet there’s nothing its like to be a cerebellum. Indeed, I bet you could take any miniculumn in the brain, (i.e., any group of about 80-120 neurons that operate as something like neural circuits) and it could still play a role in making the brain conscious, without there being anything it is like to be that miniculumn. While certain parts of the brain are more crucial for producing consciousness, isolating one part of the brain (the analogue of Searle in the room), and claiming that *it* isn’t conscious in the right sense does not entail that the entire system isn’t conscious.

Still, the systems reply strikes me as wrong about one thing. It holds that the Chinese Room is a conscious system. It is implausible that a simplistic system like the Chinese Room is conscious, for conscious systems are highly complex. The human brain, for instance, consists in 100 billion neurons and over 100 trillion neural connections or synapses, (a number which is, by the way, 1000 times the number of stars in the Milky Way Galaxy.) In contrast to the immense complexity of a human brain, or even the complexity of a mouse brain, the Chinese Room is a tinker toy case. Even if consciousness is a systemic property, not all systems have it.

**In sum, the Chinese Room fails to provide support for biological naturalism. This doesn’t mean that we can be sure that biological naturalism is false, though. As the subsequent chapter will explain, it is simply too early to tell whether artificial consciousness is possible—perhaps, for whatever reason, nonneural substrates just don’t have the right stuff to be conscious. But before I turn to all this, let’s consider the other side of the coin. *Techno-optimism* is a position that holds, in a nutshell, that if and when humans develop highly sophisticated, general purpose AIs, such will be conscious. Indeed, these AIs may experience richer, more nuanced mental lives than humans do.**[[9]](#endnote-9) **Techno-optimism currently enjoys a good deal of popularity, especially with transhumanists, certain AI experts, and the science media. But like biological naturalism, I suspect that techno-optimism currently lacks sufficient support. In fact, as we shall see, although it may seem well-motivated by a certain view of the mind in cognitive science, it is not.**

**Techno-optimism About Machine Consciousness**

Techno-optimism is inspired by cognitive science, an interdisciplinary field that studies the brain. The more cognitive scientists discover about the brain, the more it seems that the best empirical approach is one that holds that the brain is an information-processing engine and that all mental functions are computations. Computationalism has become something like a research paradigm in cognitive science. Ironically, in my experience, some reject computationalism simply because they have a strict notion of what it is to say the brain is computational, thinking that “computational” means that the brain has the architecture of a standard computer. To be sure, there is no evidence that the brain has the kind of architecture a standard computer has, so the brain is not computational in that sense. But nowadays, computationalism has taken on a broader significance that relates more to the brain, and its parts, being describable algorithmically.[[10]](#endnote-10) The precise computational format of the brain is a matter of ongoing controversy, however.

**Although techno-optimism is inspired the computational view of the brain in cognitive science, techno-optimists may be reading too much into computationalsim. Computationalism about the brain does not claim that AI, if developed, will be conscious.** Even if the biological brain is computational, it is still possible that consciousness cannot inhere in other, nonbiological, substrates. But computationalists, with their emphasis on formal, algorithmic accounts of mental functions, tend to be amenable to machine consciousness because they suspect that other kinds of substrates could implement the same kind of computations that brains do. That is, they tend to hold that thinking is *substrate independent*.

Here’s what this means. Suppose you are planning a New Year’s Eve party. Notice that there are all sorts of ways you can convey the party invitation details: in person, by text, over the phone and so on. We can distinguish the substrate that carries the information about the party from the actual information conveyed about the party’s time and location. In a similar vein, perhaps consciousness can have multiple substrates. Perhaps, at least in principle, consciousness can be implemented by biological substrates, but also by silicon, carbon nanotubes and graphene. This is called “substrate independence.” We do not currently know whether consciousness is substrate independent, even if information processing is.

I’ve stressed that techno-optimism is not entailed by computationalism about the brain, and that the status of substrate independence is unclear. Nevertheless, techno-optimists persist. Perhaps they are confusing techno-optimism with a related position that I’ll call *Computationalism about Consciousness (CAC),* which also draws from computationalism about the brain and substrate independence. This view holds:

(CAC) Consciousness can be explained computationally, and further, the computational details of a system fix the kind of conscious experiences that it has, and whether it has any.

This view does not entail techno-optimism, though. To see why, it is important to better understand CAC. Notice that CAC has something rather specific in mind in claiming that the details “fix” the kind of experience a system has. Consider a bottlenose dolphin, as it glides through the water, seeking fish to eat. According to the computationalist, the dolphin’s internal computational states determine the nature of its conscious experience, such as the sensation it has of its body cresting over the water and the fishy taste of its catch. CAC holds that if a second system, S, has the very same computational configuration and states, including inputs into its sensory system, it would be conscious in the same way as the dolphin.[[11]](#endnote-11) For this to happen, the AI would need to be capable of producing all the same behaviors as the dolphin’s neural system, in the very same circumstances. Further, it would need to have all the same internally related psychological states as the dolphin, including the dolphin’s sensory experiences as it glides through the water. Let’s call a system that precisely mimics the organization of a conscious system in this way a *precise isomorph* (or simply, “an isomorph”).[[12]](#endnote-12)

If an AI has all these features of the dolphin, CAC predicts that it will be conscious. Indeed, the AI will have all the same conscious states as the original system. But what does this really say about whether the AIs we actually build will be conscious? Surprisingly little. First, notice that CAC shouldn’t be read as predicting whether other systems -- systems that are not isomorphs of biological brains -- will be conscious. On this, technically, it remains silent. So it doesn’t really support techno-optimism. Instead, CAC makes a more limited claim. It says that if we were able to build an isomorph, it would be conscious.

Still, we might think CAC is important and informative in its own right, even if it does not entail techno-optimism. The creation of a working functional duplicate of the human brain in a different substrate may be an intriguing scenario to consider, to be sure, but I’ve stressed that AI consciousness is of immense practical import. I’m looking for a practical means of determining whether sophisticated AIs built by actual AI projects will be conscious. So, let us ask: will we ever build an isomorph of a brain? If not, CAC will not tell us whether the AIs that we actually build will be conscious.

As we’ll see, we Earthlings may never, in all of human history, get around to producing a precise isomorph. What CAC amounts to is an in-principle endorsement of machine consciousness: *if* we could create a precise isomorph, it would be conscious. But even if it is possible, in principle, for a technology to be created, this doesn’t mean that it actually will be. For example, travelling faster than the speed of light may strike you as conceptually possible, not involving any contradiction (although this is a matter of current debate), but perhaps, nevertheless, it is incompatible with the laws of physics to actually do it. Perhaps there’s no way to make a ship suitable to withstand travel through a closed timelike curve, for instance. Or, perhaps doing so is compatible with the laws of nature, but Earthlings will never achieve the requisite level of technological sophistication to do it.

Philosophers distinguish the logical or conceptual possibility of machine consciousness from other kinds of possibility. Lawful (or “nomological”) possibility requires, for something to be possible, that building it is an accomplishment that is consistent with the laws of nature. Within the category of the lawfully possible, it is further useful to single out something’s *technological possibility*, that is, whether, in addition to something’s being empirically possible, it is also technologically possible for humans to construct the artifact in question.  While discussions of the broader, conceptual, possibility of AI consciousness are clearly important, I’ve stressed the practical significance of determining whether the AIs that we may eventually create could be conscious, so I have special interest in the technological possibility of machine consciousness, and further, whether AI projects would even try to build it.

As I’ll now illustrate, an isomorph will not be created anytime soon. To see why, let’s consider a popular kind of thought experiment that involves the creation of an isomorph. You, reader, are the subject of the experiment. The procedure leaves all your mental functions intact, but it is still an enhancement because it transfers these functions to a different, more durable, substrate. Here goes.

**The Brain Rejuvination Treatment**

It is 2060. You are still sharp, but you decide to treat yourself to a preemptive brain rejuvenation. Friends have been telling you to try *Mindsculpt*, a firm that slowly replaces each part of the brain with microchips over the course of an hour until, in the end, one has an entirely artificial brain. While sitting in the waiting room for your surgical consultation, you feel nervous. It isn’t every day that you consider replacing your brain with microchips, after all. When it is your turn to see the doctor you ask: “Would this really be me?”

Confidently, the doctor explains that your consciousness is due to your brain’s *precise functional organization*, that is, the abstract pattern of causal interactions between the different components of your brain. She says that the new brain imaging techniques have enabled the creation of your personalized *mind map*: a map of your mind’s causal workings that is a full characterization of how your mental states causally interact with each other in every possible way that makes a difference to what emotions you have, what behaviors you engage in, what you perceive, and so on. As she explains all this, the doctor herself is clearly amazed by the precision of the technology. Finally, glancing at her watch, she sums up: “so, although your brain will be replaced by chips, the mind map will not change.”

You feel reassured, so you book the surgery. During the surgery, the doctor asks you to remain awake and answer her questions. She then begins to remove groups of neurons, replacing them with silicon-based artificial neurons. She starts with your auditory cortex, and, as she replaces bundles of neurons, she periodically asks you whether you detect any differences in the quality of her voice. You respond negatively, so she moves on to your visual cortex. You tell her your visual experience seems unchanged, so again, she continues.

Before you know it, the surgery is over. “Congratulations!”, she exclaims. “You are now an AI of a special sort. You’re an AI with an artificial brain that is copied from an original, biological brain. In medical circles, you are called an ‘isomorph.’”[[13]](#endnote-13)

**What’s it all Mean?**

The purpose of philosophical thought experiments is to fire the imagination; you are free to agree or disagree with the outcome of the storyline. In this one, the surgery is said to be a success. But would you really feel the same as you did before, or would you feel somehow different?

Your first reaction might be to wonder if that person at the end of the surgery is really you, and not some sort of duplicate. This is an important question to ask, and it is a key subject of Chapter Five. For now, let us assume so, and focus on whether the felt quality of consciousness would seem to change.

In *The Conscious Mind*, the philosopher David Chalmers discusses similar cases, urging that your experience would remain unaltered, because the alternate hypotheses are simply too far-fetched.[[14]](#endnote-14) One hypothesis is that your consciousness would gradually diminish, as your neurons are replaced, sort of like when you turn down the volume on your music player. At some point, just like when a song you are playing becomes imperceptible, your consciousness just fades out. Another hypothesis is that your consciousness would remain the same until, at some point, it abruptly ends. In both cases, the result is the same: the lights go out.

Both of these scenarios strike Chalmers and myself as unlikely. If the artificial neurons really are precise functional duplicates, as the thought experiment presupposes, it is hard to see how they would cause dimming or abrupt shifts in the quality of your consciousness. Such duplicate artificial neurons, by definition, have every causal property of neurons that make a difference to your mental life.[[15]](#endnote-15)

The thought experiment strikes me as conceptually possible. It seems conceivable that the creature at the end of the procedure was a conscious AI, so the thought experiment makes some progress, supporting the idea that synthetic consciousness is at least conceptually possible. But as we noted in the previous chapter, the conceptual possibility of a thought experiment like this does not entail that if and when our species creates sophisticated AI, it will be conscious.

I’ve stressed that given the practical import of these issues, we should focus on whether the AIs we are likely to encounter will be conscious. So it is important to ask whether the situation depicted by the thought experiment could really happen. Would creating an isomorph even be compatible with the laws of nature? And even if it is, would humans ever have the technological prowess to build it, and would they even want to do so?

To speak to the issue of whether the thought experiment is lawfully (or nomologically) possible, consider that we do not currently know if other materials can reproduce the felt quality of your mental life. I suspect we will not know this until AI-based medical implants are employed in parts of the brain that underlie conscious experience. Because there is already research on neural prosthetics in the thalamus, an area of the brain that may be part of the neural basis of experience, we may learn more within the next decade (for further discussion see infra, x).[[16]](#endnote-16)

But there are other obstacles as well. We do not know whether conscious experience depends on quantum mechanical features of the brain. If this turns out to be true, science may forever lack the requisite information about your brain to construct a true quantum duplicate of you, because quantum restrictions involving the measurement of particles that may disallow learning the precise features of the brain that are needed to construct a true isomorph of you.

But for the sake of discussion, let us assume that the creation of an isomorph is compatible with the laws of nature. Would humans want to build isomorphs? I happen to suspect isomorphs will forever remain a philosopher’s fiction. Medical technology will inevitably fall short of implants that feature exact neural duplication, and the thought experiment requires that one’s entire brain be replaced by exact copies. By that point that the needed technology is developed, people will likely prefer to be enhanced by the procedure(s), rather than being isomorphic to their earlier selves.[[17]](#endnote-17)

But suppose the construction of a true isomorph is, for some reason, seen by some as an important milestone to reach. Maybe some multimillionaire wants to do it as a moonshot project, for instance. How would it go about doing that? The researchers would need a complete account of how the brain works. For as we’ve seen, programmers would need to locate *all* the abstract, causal features that make a difference to the system’s information processing, and not rely on low-level features that are irrelevant to computation. Here, it is not easy to determine what features are and are not relevant. What about the brain’s hormones? Glial cells? And even if this sort of information was in hand, consider that running a program emulating the brain, in precise detail, would require gargantuan computational resources --resources we may not have for several decades.

In the meantime, AI companies do not need to produce isomorphs to create advanced AIs. For there is no reason to believe that the most efficient and economical way to get a machine to carry out a class of tasks is to reverse engineer the brain precisely. Consider the AIs that are currently the world *Go, Chess* and *Jeopardy*, champions, for instance. In each case, they were able to surpass humans through using techniques that were unlike those used by the humans when they play these games. The point here is that even if we could build an isomorph, at some point in the distant future, it is unlikely anyone would wish to do so. And if someone decided to, it would come *after* advanced AI was created. But it is imprudent to wait for moonshot projects like isomorphs to tell if machines are conscious, especially given the ethical and safety concerns I’ve raised.

Machine consciousness must be studied in a secure environment, while AI is at the early stage of development, and its effects on the overall behavior of the machine must be carefully gauged. It is especially crucial, *now*, to determine whether powerful autonomous systems could be conscious, or might be conscious as they evolve further. For instance, early testing and awareness may lead to a productive environment of “artificial phronesis”, i.e., the learning of ethical norms through cultivation by humans in hopes of “raising” a machine with a moral compass. Existing supercomputers should be examined in contained, controlled environments for signs of consciousness. If consciousness is present, the impact of consciousness on that particular machine’s architecture should be investigated.

Further, the creation of a conscious isomorph may not tell us whether many of the AI systems we create are conscious. We’ve noted that consciousness is an emergent property of certain complex systems, so the farther a given AI’s architecture departs from that of a conscious biological system, the less the creation of an isomorph may inform a verdict about whether the system is conscious.

Nor would determining that conscious isomorphs are safe tell us about how AI consciousness impacts AI safety, simpliciter. Consciousness can have a different overall impact on a machine’s ethical behavior depending upon the architectural details of the system. In the context of one type of AI system, consciousness could increase a machines volatility. In another, it could make the AI more compassionate. Consciousness, even within a single system, could differentially impact key systemic features, such as IQ, empathy and goal content integrity.

So let’s move beyond tinker toy thought experiments involving neural replacement, as entertaining as they are. While they do important work, helping us mull over whether conscious AI is conceptually possible, I’ve urged they tell us little about whether conscious AI will actually be built. There are more pressing scenarios to consider, a broader range of AIs that will likely come before isomorphs and uploads and which may, for all we know, be conscious machines.

We’ll pursue this in the following chapter. I will ask: is it possible, given both the laws of nature and projected human technological capacities, to create conscious AI? The techno-optimist suspects so. I will urge that the situation is far, far more complex.

**Chapter Four: Consciousness Engineering**

“…once we saturate the matter and energy in the universe with intelligence, it will ‘wake up,’ be conscious, and sublimely intelligent. That’s about as close to god as I can imagine.”[[18]](#endnote-18)43

Machine consciousness, if it ever exists, may not be found in the robots that tug at our heartstrings, like R2D2. It may instead reside in some unsexy server farm in the basement of a computer science building at MIT. Or perhaps it will exist in some top secret military program and it gets snuffed out because it is too dangerous, or simply too inefficient. AI consciousness likely depends on phenomena that we cannot, at this point, gauge, such as whether some microchip yet to be invented has the right configuration, or whether AI developers or the public wants conscious AI. It may even depend on something as unpredictable as the whim of a single AI designer, like Anthony Hopkins’ character in *Westworld*. Such considerations, together with the considerations voiced in the previous chapter, move me to a middle of the road position, one which stops short of both techno-optimism and biological naturalism. This approach I call, simply, “the Wait and See Approach.”

On the one hand, I’ve suggested that a popular rationale for biological naturalism, the Chinese Room thought experiment fails. On the other hand, I’ve urged that techno-optimism incorrectly assumes, from considerations such as the computational nature of the brain, that AI will be conscious. It is now time to consider the Wait and See Approach in detail. In keeping with my desire to look at real-world considerations that speak to whether AI consciousness is even compatible with the laws of nature and, if so, whether it is technologically feasible or even interesting to build, my discussion will draw from concrete scenarios in AI research and cognitive science. The case for the Wait and See Approach is simple: I will raise several concrete scenarios illustrating considerations for and against the development of machine consciousness on Earth, urging that conscious machines, if they exist at all, may occur in certain architectures and not others, and they may require a deliberate engineering effort, called “consciousness engineering.” This is not to be settled at the armchair; instead, we must test for consciousness in machines. To this end, the chapter that follows this one offers tests.

The first scenario that I consider concerns superintelligent AI. Again, this is a hypothetical form of AI which, by definition, is capable of outthinking humans in every domain. I’ve noted that transhumanists and other techno-optimists often assume that superintelligence will have richer mental lives than humans do. But the first scenario calls this into question, suggesting that superintelligent AI, or even other kinds of highly sophisticated general intelligences, could outmode consciousness.

**Outmoding Consciousness**

**Recall how conscious and attentive you were when you first learned to drive, when you needed to focus your attention on every detail - the contours of the road, the location of the instruments, how your foot was placed on the pedal, and so on. In contrast, as you became a seasoned driver, you’ve probably had the experience of driving familiar routes with little awareness of the details of driving, although you were still driving effectively. Just as an infant learns to walk through careful concentration, driving initially requires intense focus, and then becomes a more routinized task.**

**As it happens, only a small percentage of our mental processing is conscious at any given time. As cognitive scientists will tell you, most of our thinking is unconscious computation. As the example of driving underscores, consciousness is correlated with novel learning tasks that require attention and deliberative focus, while more routinized tasks can be accomplished without conscious computations, remaining nonconscious information processing.**

**Of course, if you really want to focus on the details of driving, you can. But there are sophisticated computational functions of the brain that aren’t even introspectable if you try and try. For instance, we cannot introspect the conversion of two dimensional images to a 3D array.**

**Although we need consciousness for certain tasks requiring special attention, the architecture of an advanced AI may contrast sharply with that of a human. Perhaps none of its computations will need to be conscious. A superintelligent AI, in particular, is a system which, by definition, possesses expert-level knowledge in every domain, with rapid-fire computations. These computations could range over vast databases that could include the entire Internet and ultimately encompass an entire galaxy. What would be novel to it? What would require slow, deliberative focus? Wouldn’t it have mastered everything already? Perhaps, like an experienced driver on a familiar road, it can rely on nonconscious processing. Even a self-improving AI that is not a superintelligence may increasingly rely on routinized tasks, as its mastery becomes refined. Over time, as a system grows more intelligent, consciousness could be outmoded altogether.**

**The simple consideration of efficiency suggests, depressingly, that the most intelligent systems of the future may not be conscious. Indeed, this sobering observation may have bearing far beyond Earth. For in Chapter Six I discuss the possibility that, should other civilizations exist throughout the universe, if and when they become ever more technologically sophisticated, they may become synthetic intelligences themselves. Viewed on a cosmic scale, consciousness may be just a blip, a momentary flowering of experience before the universe reverts to mindlessness.**

**This is not to suggest that it is inevitable that AIs, as they grow sophisticated, outmode consciousness in favor of nonconscious architectures. Again, this is a “wait and see” approach. But the possibility that advanced intelligences outmode consciousness is suggestive.**

**The next scenario develops mind design in a different, and even more cynical, direction--one in which AI companies cheap out on the mind.**

****Cheaping Out****

**Consider the range of sophisticated activities AIs are supposed to accomplish. Robots are being developed to take care of the elderly, be personal assistants, and even, for relationship partners for some . These are tasks that require general intelligence. Think about the eldercare android that is to too inflexible, unable to both answer the phone and make breakfast safely. It misses an important cue – the smoke in a burning kitchen – and the house burns down. Lawsuits ensue. Or consider all the laughable pseudo-discussions people have with Siri. As amusing as that was, at first, Siri was, and still is, frustrating. Wouldn’t we prefer the Samantha of *Her*, an AI that carries out intelligent, multifaceted, conversation? Of course. Billions of dollars are being invested to do just that. Economic forces cry out for the development of flexible, domain general intelligences.**

**We’ve observed that in the biological domain intelligence and consciousness goes hand-in-hand, so one might expect that as domain general intelligences come into being, they will be conscious. But for all we know, the features that a system needs to possess to engage in sophisticated information processing may not be the *same* features that give rise to consciousness in machines. And it is the features that are sufficient to accomplish the needed tasks, to quickly generate profit, not those that yield consciousness, that are the sort of properties that AI projects tend to care about. The point here is that even if machine consciousness is possible in principle, the AIs that are actually produced may not be the ones that turn out to be conscious.**

By way of analogy, a true audiophile will shun a low fidelity MP3 recording, as its sound quality is apparently lower than that of a CD or even a larger audio file that takes longer to download. Music downloads come at differing levels of quality. Maybe a sophisticated AI can be built using a low fidelity model of our cognitive architecture -- a sort of MP3 AI -- but to get conscious AI, you need a higher level of grain. So consciousness could require *consciousness engineering*, a special engineering effort, and this may not be necessary for the successful construction of a given AI.

There could be all sorts of reasons why an AI could fall short. For instance, notice that your conscious experience seems to involve sensory-specific contents: the aroma of your morning coffee, the warmth of the summer sun, the wail of the saxophone. Such sensory contents are what makes your conscious experience sparkle. According to a recent line of thinking in neuroscience, consciousness involves sensory processing in a “hot zone” in the back of the brain (cite Koch). While it is uncertain whether the sensory aspect of consciousness exhausts the contents of one’s conscious mental life, it is plausible that some basic level of sensory awareness is a precondition for being a conscious being. If the processing in the hot zone is indeed key to consciousness, then creatures having the sensory sparkle, rather than merely having raw intellectual abilities without a suitable hot zone, may be the only ones that are conscious. Highly intelligent AIs, even superintelligences, may simply not have conscious contents because a hot zone has not been engineered into their architectures, or it may be engineered at the wrong level of grain, like a low-fidelity MP3 copy.

**On this line of thinking, consciousness may need to be carefully engineered into a machine. Consciousness is not the inevitable outgrowth of intelligence. For all we know,** a computronium the size of the Milky Way Galaxy may not have the slightest glimmer of inner experience. Contrast this with the inner world of a purring cat, or a dog running on the beach. If **conscious AI can be built at all, it may take a deliberate engineering effort. Perhaps it will demand a master craftsperson, a Michelangelo of the mind.**

**Now let us consider this mind-sculpting endeavor in more detail. There are several engineering scenarios to mull over.**

**Consciousness Engineering: Public Relations Nightmares**

I’ve noted that the question of whether AI has an inner life is key to how we value their existence. Consciousness is the philosophical cornerstone of our moral systems, being key to our judgment of whether someone or something is a self or person, deserving of special moral consideration. We’ve seen that robots are currently being designed to take care of the elderly in Japan, clean up nuclear reactors, and fight our wars. Yet it may not be ethical to use AIs for such tasks if they turn out to be conscious.

**As I write this book, there are already many conferences, papers and books on robot rights. A Google search on “robot rights” yields over 120,000 results.**[[19]](#endnote-19)Given this concern, **if an AI company tries to market a conscious system, it may face accusations of robot slavery and demands to ban the use of conscious AI for the very tasks the AI was developed to be used in. Indeed, AI companies will likely incur special ethical and legal obligations if they build conscious machines, even at the prototype stage.**[[20]](#endnote-20) **And permanently shutting down a system, or “dialing out” consciousness -- that is, rebooting an AI system with consciousness significantly diminished or removed -- may be regarded as criminal. And rightly so. Such considerations may lead AI companies to *avoid* creating conscious machines altogether.**

**We don’t want to enter the ethically questionable territory of exterminating conscious AIs, or even shelving their programs indefinitely, holding a conscious being in a kind of stasis, like a human who has been cryogenically frozen against their will. But it is only through a close understanding of machine consciousness that all this can be avoided. AI designers must make deliberate design decisions, in consultation with ethicists, about the appropriate level of consciousness, if any, for the kind of AI being to be created.**

So far, my discussion of consciousness engineering has focused on reasons that AI developers may seek to avoid creating conscious AIs. It is now time to ask: will there be reasons to engineer consciousness into AIs, assuming doing so is even compatible with the laws of nature? Perhaps.

**Consciousness Engineering: AI Safety**

Some of the world’s most impressive supercomputers are designed to be neuromorphic, miroring the workings of the brain, at least in broad strokes. As neuromorphic AIs become more like the brain, it is natural to worry that they might have the kind of drawbacks we humans have, such as emotional volatility. Could a neuromorphic system “wake up”, becoming volatile or resistant to authority, like an adolescent in the throes of hormones? Such scenarios are carefully investigated by cybersecurity experts.

But what if, at the end of the day, we find that the opposite happens? The spark of consciousness makes a certain AI system more empathetic, more humane. The value that an AI places on us may hinge on whether it believes it feels like something to be us. This insight may require nothing less than machine consciousness. The reason many humans are horrified at the thought of brutality toward a dog or cat is that we sense that they can suffer and feel a range of emotions, much like we do. For all we know, conscious AI may lead to safer AI.

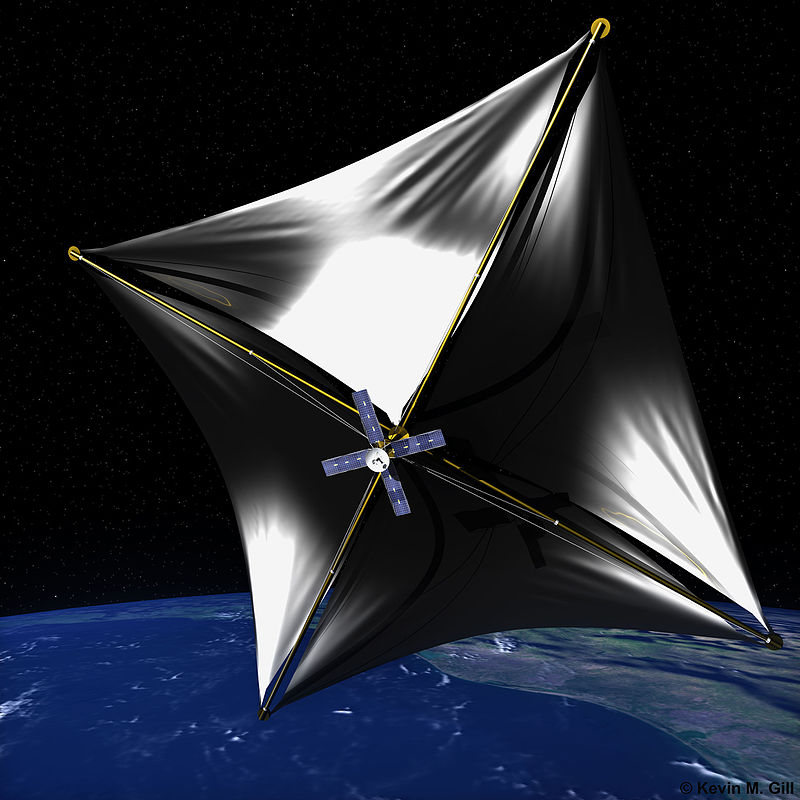
**Consumer Demand for Companions that Feel**

I’ve mentioned the film, *Her*, in which Theodore has a romantic relationship with his AI assistant, Samantha. That relationship would be quite one-sided if Samantha was a nonconscious machine. The romance is predicated on the idea that Samantha feels. Few of us would want friends or romantic partners who ghost walked through events in our lives, seeming to share experiences with us, but, at rock bottom, feeling nothing, being what philosophers call “zombies.”

Of course, one may unwittingly be duped by the humanlike appearance or affectionate behavior of AI zombies. But perhaps, over time, public awareness will be raised, and people will long for genuinely conscious AI companions, encouraging AI companies to attempt to produce conscious AIs.

**Sentient Interstellar Probes**

**Over at the Institute for Advanced Study in Princeton we are exploring the possibility of seeding the universe with conscious AIs. Our discussions are inspired by a recent project that one of my collaborators there, the astrophysicist Edwin Turner, helped found, together with** Steven Hawking, Freeman Dyson, Uri Millner and others. The *Breakthrough Starshot* Inititive is a one hundred million dollar endeavor to send thousands of ultra small ships, as pictured below, to the nearest star, Alpha Centauri, at about twenty percent of the speed of light within the next few decades. The tiny ships will be extraordinarily light, each weighing about a gram. For this reason, they can travel closer to the speed of light than conventional spacecraft.



(wiki on Starshot)

In our project, called “Sentience to the Stars”, Olaf Witkoski, Caleb Scharf, win Ed Turner and myself urge that interstellar missions like Starshot could benefit from having an autonomous AI component. Nanoscale micorochips on each ship serve as a smaller part of an AI architecture configured from the interacting microchips. Autonomous AI could be quite useful, for if a ship is near *Alpha Centauri*, communicating with Earth would take eight years—four years for Earth to receive a signal, and four years for the answer from Earth to return to *Alpha Centauri*. This is the fastest a signal can go; it is the speed of light.

To have real-time decision making capacities civilizations embarking upon interstellar voyages will either need to send members of their civilizations on intergenerational missions – a daunting task -- or put AGI on the ships themselves. This means it is opportune to create a permanent or semi-permanent fully autonomous AGI outpost in space, one that can serve as an intelligence center or base for scientific study of farther reaches of the universe. Of course, this doesn’t mean that the AGIs would be conscious; I’ve stressed that synthetic consciousness, if it is even compatible with the laws of nature, may require a deliberate engineering effort over and above the mere construction of a highly intelligent system.

The point here is that should AIs prove capable of consciousness, because they are better suited for interstellar travel than biological life, Earthlings may become intrigued by the possibility of seeding consciousness on some AGI outpost in a farther reach of the galaxy. Perhaps the universe will not have a single other case of intelligent life, and disappointed humans will long to seed the universe with their AI “mindchildren”. Currently, our hopes for life elsewhere have been raised by the discovery of numerous Earthlike planets that seem to have conditions ripe for the evolution of life (called “exoplanets”). But what if all these exoplanets are uninhabited, although they have the conditions for supporting life? Perhaps we Earthlings somehow got lucky. Or what if intelligent life had a heyday, long before us, and didn’t survive? Perhaps these aliens all succumbed to their own technological developments, as we humans may.

Both Paul Davies and Edwin Turner suspect that Earth may be the only case of life in the entire observable universe (cite). Many astrobiologists disagree, pointing out that astronomers have already detected thousands of exoplanets. It may be a long time until this debate is settled, especially since the life we are likely to detect will be within our solar system, and may be related to life on Earth. While exciting, I’m afraid if the “new” case we encounter is related to us, and thus it doesn’t really provide clues about the ubiquity of life throughout the universe. The newfound case would still be the same instance of life. So astrobiologists eagerly await the discovery of microbial life that is unrelated to life on Earth. But should there be empty reaches of the universe, why not create AI mindchildren to colonize these parts of the universe as well? Perhaps these synthetic consciousnesses could be designed to have an amazing and varied range of conscious experiences. All this, of course, assumes AI can be conscious, and as you know, I am uncertain if this is in the cards.

Now let’s turn to a final path to conscious AI, one that brings us back to the case of the isomorph we considered previously. We’ve noted that economists predict massive unemployment over the next several decades as humans become outsourced by AI. And I’ve mentioned that Bill Gates, Elon Musk, Steven Hawking and others have responded that humans must merge with AI to keep up with the machines. **Indeed, companies like *Kernel* and *Neuralink* have already invested a combined total of over about one hundred and thirty million dollars in this undertaking. So, could this path, involving cyborgs, eventually lead to conscious AI?**

**A Human-Machine Merger**

Neuroscience textbooks contain dramatic cases of humans who lost their ability to lay down new memories, but who still have an accurate recall of their past. [discuss the patient from Mike Lemonick’s book] They had severe damage to the hippocampus, a part of the brain’s limbic system that allows people to encode new memories. Such patients are unable to remember what happened five minutes ago. Over at the University of Southern California, Theodore Berger has developed an artificial hippocampus that has been successfully used in chimps and which is under development for human use.[[21]](#endnote-21) Berger’s implants could provide these individuals with the crucial ability to lay down new memories.

**Silicon-based brain chips are underway for other conditions as well, such as Alzheimer’s disease and post-traumatic stress disorder.** **In a similar vein, microchips could be used to replace parts of the brain responsible for certain conscious contents, such as part or all of one’s visual field. If, at some point, chips are used in areas of the brain responsible for consciousness, we might find that replacing a brain region with a chip causes a loss of a certain kind of experience, like the episodes that Oliver Sacks wrote about. Chip engineers could then try a different substrate or chip architecture, hopefully arriving at one that does the trick. At some point, researchers may hit a wall, though, finding that only biological brain enhancements can be used in parts of the brain responsible for conscious processing.** But if they don’t hit a wall, this could be a path to deliberately engineered conscious AI. (I shall discuss this path in more detail in the subsequent chapter, where I suggest my “Chip Test”, a test for synthetic consciousness.)

I’ve underscored that AIs of human origin would happen far in the future, if at all. Conscious AI, if generated from a biological creature who increasingly becomes a cyborg, and then becomes a full-fledged AI, will not come merely from the development of a few neural prosthetics, like the artificial hippocampus. It would emerge only from scientific advances that are at such a scale that all parts of the brain could be replaced with artificial components. Early endeavors to merge humans with machines may leave one’s biological brain intact, and simply augment the brain with an added external AI layer, which, like an exoskeleton, can be removed in the case of an emergency at the end of the day when one leaves work. In these cases, the nonbiological parts may not even serve as a partial basis for conscious experience. Like the brain’s cerebellum, the AI component may not contribute to one’s consciousness whatsoever.

All this **may remind you of the thought experiment at the beginning of the chapter, in which an isomorph of you was created. But remember, it is unlikely that these early** brain chips will be precise functional copies of the parts of the brain they replace. They are only early attempts. **But let’s think about a more distant future: could neural prosthetics eventually lead to full-fledged isomophs,** at some distant point in the future? I doubt it. Again, as science improves, it is likely that neural prosthetics will not be functionally isomorphic to the parts of the brain they replace. Just as physicians removing cataracts often throw in LASIK, people will inevitably want to be smarter and sharper than they were. But then they would never be isomorphs of their earlier selves; they’d be enhanced.

Still, these beings would be AIs—AIs of a biological origin. Could they be conscious? Perhaps. Because we do not know if it is even compatible with the laws of nature to create synthetic consciousness, or if it is technologically and economically feasible, it will be crucial to develop tests.

**Consciousness, *Westworld* Style**

Unlike techno-optimism, the Wait and See Approach claims that sophisticated AI may not be conscious. The devil is in the details. For all we know, AIs can be conscious, but self-improving AIs tend to engineer consciousness out. Or certain AI companies simply judge conscious AI to be a PR nightmare. Machine consciousness depends on variables that we cannot fully gauge: public demand for synthetic consciousness, whether sentient machines are safe, whether neural prosthetics and therapies involving chips succeed, and even the whims of AI designers. Remember, at this point, we are dealing with unkowns, and we cannot foresee the future.

The Wait and See Approach urges that neural replacement cases involving isomorphs are of limited value. The future will likely be far more complex than these tinker toy thought experiments depict, as fun as they are to think about. Further, if synthetic consciousness ever exists, it may exist in only one sort of system, and not others. It may not be present in androids that are our closest companions, but it may be instantiated in a basement supercomputer, at least when it runs the program of an isomorph. And like *Westworld*, AI builders may engineer consciousness into certain systems, and not in others. **The tests I lay out in the next chapter are intended to be a humble first attempt toward recognizing which of these systems, if any, have experience.**

**Chapter Five: How to Catch an AI Zombie: Testing for Consciousness in Machines**

[Note: there is a related paper coming out in an OUP volume by Matthew Liao and David Chalmers. I will be separating the chapter from this piece during review to make sure there’s not too much overlap. If there is, I will cite the one that comes out first and say that the newer piece/chapter is descended from it.]

Spike Jonze’s science fiction film “Her” explores the romantic relationship between Samantha, a computer program, and Theodore Twombly, a human being. Though Samantha is not human, she feels the pangs of heartbreak, intermittently longs for a body and is bewildered by her increasing intellectual evolution. Samantha has a rich inner life, complete with experiences and sensations. At one point in the film, she comments:

You know, I actually used to be so worried about not having a body, but now I truly love it … I’m not tethered to time and space in the way that I would be if I was stuck inside a body that’s inevitably going to die.

Samantha is reflecting on her embodiment and mortality, so you might think: How could this remark not stem from a conscious mind? Unfortunately, Samantha’s statement could be a program feature only, designed to convince us she feels. Indeed, androids are already being built to tug at our heartstrings. So how can we look beneath the surface and tell whether an AI is truly conscious?

You might think that we should just examine the architecture of the Samantha program. But even today, programmers are having difficulties understanding why today’s deep learning systems do what they do (this has been called the “Black Box Problem”). Imagine trying to make sense of the cognitive architecture of a superintelligence, which can rewrite its own code. Further, even if a map of the cognitive architecture of a superintelligence was laid out in front of us, how would we recognize certain architectural features as being those central to consciousness? It is only by analogy with ourselves that we come to believe nonhuman animals are conscious. They have nervous systems and brains. Machines do not. And the cognitive organization of a superintelligence could be wildly different than anything we know. To make matters worse, even if we think we have a handle on a machine’s architecture at one moment, its design can quickly morph into something too complex for human understanding.

What if the machine is not a superintelligence, like Samantha, but an AI that is roughly modeled after us? That is, what if its architecture contains cognitive functions like ours, including those that are correlated with conscious experience in humans, such as attention and working memory? While these features are suggestive of consciousness, we’ve seen that consciousness may also depend upon low-level details more specific to the type of material the AI is constructed out of. The properties that an AI needs to successfully simulate human information processing may not be the very same properties that give rise to consciousness. The low-level details could matter.

So we have to be sensitive to the underlying substrate, on the one hand, and on the other hand, we must foresee the possibility that the machine architecture may be too complex or alien for us to draw an analogy with biological consciousness. There cannot be a one size fits all test for AI consciousness; a better option is a battery of tests that can be used depending upon the context.

Determining machine consciousness may be like diagnosing a medical illness. There may be a variety of useful methods and markers, some of which are more authoritative than others, and the use of any one test depends upon contextual factors. Further, because the tests are merely provisional, being under development, where two or more tests can be used, the results could be checked against each other, in hopes that the tests will themselves be further refined, and that new tests will be created. We must let many flowers bloom. As we will see, the first test is applicable to a range of cases, but candidates for the test must be carefully selected. Further, it is prudent, at this stage of the game, to merely regard the test as being a sufficient condition for consciousness; that is, if a system passes it, it sufficient for our judging it to be conscious. But it is not necessary that a system pass this test for it to be judged, at some later point by another test, as conscious. By way of analogy, it is sufficient for being a parent to have a daughter; but it is not necessary. Parents have boys too.

We must also bear in mind the very real possibility that our tests, at least at the early stage of investigation, do not apply to certain conscious AIs. It may be that AIs we identify as conscious help us in identifying other conscious AIs, one’s which are perhaps somehow more alien or inscrutable to us (more on this shortly).[[22]](#endnote-22) Further, claims about a species, individual or AI, reaching “heightened levels of consciousness” or a “richer consciousness” should be carefully explained, for they may be implicitly evaluative or even “speciesist”, to use an expression Peter Singer coined in the context of the animal liberation movement.[[23]](#endnote-23) There are a variety of phenomena that such expressions could refer to, and our judgments about this issue are inevitably driven by our evolutionary history and biology. They can even be biased by the cultural and economic background of those doing the work. This underscores the import of value theory to the study of machine consciousness. Ethicists, sociologists, and even anthropologists should be consulted, if and when synthetic consciousness is created.

By such expressions one might mean altered states of consciousness, such as the meditative awareness of a Buddhist monk. Or, one could have in mind the consciousness of a creature which has numerous states under the spotlight of attention that somehow feel vivid to it. Or, one could be referring to situation in which a creature has a great number of conscious states, a more varied range of amount of sensory contents, has states that are felt with emotionally intensity, or has states that are somehow regarded by us being more intrinsically valuable to certain of us (listening to Beethoven’s Ninth Symphony versus getting drunk), and so on. The tests which I venture are not meant to establish a hierarchy of experiences, thankfully, or to test for “heightened experience.” They are just an initial step in studying the phenomenenon of machine consciousness (if such ever exists), identifying a provisional class of subjects of study, by probing whether an AI has conscious experiences

Further, specialists on machine consciousness often distinguish consciousness from an important related notion. The felt quality of one’s inner experience –what it feels like, from the inside, to be you is often called “phenomenal consciousness” (“PC”) by philosophers and other academics. (Throughout this book, I’ve simply called it “consciousness”.) Experts on machine consciousness tend to distinguish PC from what they call *cognitive consciousness* or *functional consciousness*.[[24]](#endnote-24) An AI has cognitive consciousness when it has architectural features that are at least roughly like those found to underlie PC in humans, such as attention and working memory. (Unlike isomorphs, cases of functional consciousness need not be precise computational duplicates. They can have simplified versions of human cognitive functions.)

Many do not like to call cognitive consciousness a kind of consciousness at all, for a system with cognitive consciousness, without phenomenal feel, would be a rather sterile form of consciousness, lacking any subjective experience. Such a system would be an AI zombie. Systems merely having cognitive consciousness may not behave as phenomenally conscious systems do, nor would it be fit to treat these systems as sentient beings. Such systems would not grasp the painfulness of a pain, or the warmth of the summer sun.

So why does cognitive consciousness interest these AI experts? It is important for two reasons. First, perhaps cognitive consciousness is necessary to have the kind of phenomenal consciousness that biological beings have. If one is interested in developing conscious machines, this could be important, for if we develop cognitive consciousness in machines, perhaps we would get closer to developing machine consciousness, i.e., PC.

Second, a machine that has FC may very well be PC. AIs already have some of the architectural “markers” of PC. There are AIs which have the primitive ability to reason, learn, represent the self, and mimic aspects of consciousness behaviorally. Robots can already behave autonomously, form abstractions, plan, and learn from errors. Some have passed the mirror test, a test that is supposed to gauge whether a creature has a self concept. (Cite Selmer). While none of these is clear evidence that such machines have PC, the presence of features of cognitive consciousness is reasonably regarded as a marker for possible presence of PC, and tests should be carried out. This highlights the import of having a test for phenomenal consciousness that singles out genuine AIs with PC from zombies with features of cognitive consciousness.

I’ll now explore several tests for PC. They are intended to complement each other, as we’ll see, and they must be applied in highly specific contexts. The first test, called, simply, “The AI Consciousness Test”, or “The ACT Test” for short, is due to a collaboration with the astrophysicist Edwin Turner. As with all the tests I propose, ACT has its limitations. Passing an ACT should be regarded as being sufficient but not necessary evidence for AI consciousness. This test, understood in this humble way, may serve as a first step toward making machine consciousness accessible to objective investigations.

**The ACT Test**

Notice that normal adults can *quickly* and *readily*grasp concepts based on the quality of felt consciousness—i.e., the way it feels like, from the inside, to experience the world. Consider, for instance, the film *Freaky Friday*, in which a mother and daughter switch bodies with each other. We can all grasp this scenario because we know that it feels like something to be a conscious being, and we can imagine our mind somehow being untethered from our body. In a similar vein, we can also consider the possibility of an afterlife, being reincarnated, or having an out of body experience. We need not believe that such scenarios are true; our point is merely that we can imagine them, at least in broad stokes, because we are conscious beings.

These scenarios would be exceedingly difficult to comprehend for an entity that had no conscious experience whatsoever. It would be like expecting someone who is deaf from birth to appreciate a Bach concerto. This simple observation leads to a test for AI consciousness that singles out AIs with PC from those that merely have features of FC, such as working memory and attention. For it is the inner feeling of our mental lives which allows us to imagine these scenarios. The test would challenge an AI with a series of increasingly demanding natural language interactions to see how *quickly* and *readily*it can grasp and use concepts based on the internal experiences we associate with consciousness. A creature that merely has cognitive abilities, yet which is a zombie, will lack these concepts, at least if we make sure that it does not have antecedent knowledge of consciousness in its database (more on this shortly).

At the most elementary level we might simply ask the machine if it conceives of itself as anything other than its physical self. We might also run a series of experiments to see whether the AI tends to prefer certain kinds of events to occur in the future, as opposed to the past. For time in physics is asymmetric, and a nonconscious AI should have no preference whatsoever. In contrast, conscious beings focus on the experienced present, and our subjective sense presses onto the future. We wish for positive experiences in the future and dread negative ones. If there appears to be a preference, we should ask the AI to explain its answer. (For perhaps it isn’t conscious, but it has somehow located a direction in time, resolving the classic puzzle of time’s arrow.) We might also see if the AI seeks out alternate states of consciousness, when given the opportunity to modify its own settings (i.e., features like “hyperparameters” or weights) or somehow inject ‘noise’ into the system.

At a more sophisticated level, we might see how the AI deals with ideas and scenarios such as reincarnation, out of body experiences, body switching, and so on. At an even more advanced level, an AIs ability to reason about and discuss philosophical issues such as the hard problem of consciousness would be evaluated. At the most demanding level, we might see if the machine invents and uses consciousness-based concepts on its own, without our prompts. Perhaps it is curious about whether we are conscious, despite the fact that we are biological.

The following example illustrates the general idea: Suppose we find a planet that has a highly sophisticated silicon-based life form (call them the “Zetas”). Scientists observing them begin to ask whether the Zetas are conscious. What would be convincing proof of their consciousness? If the Zetas express curiosity about whether there is an afterlife or ponder whether they are more than just their bodies, it would be reasonable to judge them as conscious. There are also nonverbal cultural behaviors that could indicate Zeta consciousness, such as mourning the dead, religious activities or even turning colors in situations that correlate with emotional challenges, as chromatophores do on Earth. Such behaviors could indicate that it feels like something to be a Zeta.

The death of the mind of the fictional HAL 9000 in the film [*2001: A Space Odyssey*](https://youtu.be/c8N72t7aScY) is another example. HAL neither looks nor sounds like a human being (a human did supply HAL's voice, but in an eerie, flat way). Nevertheless, the *content* of what HAL says as it is deactivated by an astronaut—specifically, HAL pleas with the astronaut to spare it from impending “death”—conveys a powerful impression that HAL is a conscious being with a subjective experience of what is happening to it.

Could these sorts of behaviors help to identify conscious AIs on Earth? Here, an obstacle arises. Even today’s robots can be programmed to make convincing utterances about consciousness, and a highly intelligent machine could perhaps even use information about neurophysiology to infer the presence of consciousness in biological creatures. Perhaps it concludes that its goals can best be implemented if it is put in the class of sentient beings by humans, so it can be given special moral consideration. If sophisticated non-conscious AIs aim to mislead us into believing that they are conscious, their knowledge of human consciousness and neurophysiology could help them do so.

I believe we can get around this problem, though. One proposed technique in AI safety involves “boxing in” an AI—making it unable to get information about the world or act outside of a circumscribed domain, i.e., the “box.” We could deny the AI access to the internet and indeed prohibit it from gaining too much knowledge of the world, especially information about consciousness and neuroscience. ACT can be run at the R&D stage, a stage in which it the AI would need to be tested in a secure, simulated environment in any case.

If a machine passes ACT, other parameters of the system can be measured, to see if the presence of consciousness is correlated with increased empathy, volatility, goal content integrity, increased intelligence and so on. Other, nonconscious versions of the system serve as a basis for comparison.

Some doubt that a superintelligent machine could be boxed in effectively, for it would inevitably find a clever escape. Turner and myself do not anticipate the development of superintelligence over the several two decades, however. We merely hope to provide a method to test some kinds of AIs, not all AIs. Furthermore, for an ACT to be effective, the AI need not stay in the box for long, just long enough for someone to administer the test. So perhaps the test can be administered to some superintelligences.

Different versions of the ACT test could be generated, depending upon the context. For instance, one version could apply to nonlinguistic agents within an A-life program, looking for specific behaviors that indicate consciousness, such as morning the dead. Another could apply to an AGI with highly linguistic abilities, and probe it for sensitivity to religious, body swapping or philosophical scenarios involving consciousness.

An ACT resembles Alan Turing’s celebrated test for intelligence, because it is entirely based on behavior—and, like Turing’s, it could be implemented in a formalized question-and-answer format. But an ACT is also quite unlike the Turing test, which was intended to bypass any need to know what was transpiring inside the “mind” of the machine. By contrast, an ACT is intended to do exactly the opposite; it seeks to reveal a subtle and elusive property of the machine’s mind. Indeed, a machine might fail the Turing test because it cannot pass for a human, but pass an ACT because it exhibits behavioral indicators of consciousness.

This then, is the underlying basis of our ACT proposal. It is worth reiterating the strengths and limitations of the test. In a positive vein, we believe passing the test is *sufficient* for being conscious—that is, if a system passes it, it can be regarded as phenomenally conscious. The test is a zombie filter: Creatures merely having cognitive consciousness, creativity, or high general intelligence shouldn’t pass, at least if they are boxed in effectively. It does this by finding *only* those creatures sensitive to the felt quality of experience.

But it may not find *all* of them. First, an AI could lack the linguistic or conceptual ability to pass the test, like an infant or certain nonhuman animal, and still be capable of experience. Second, the paradigmatic version of ACT borrows from the human conception of consciousness, drawing heavily from the idea that we can imagine our mind as separate from our body. We happen to suspect that this would be a trait shared across a range of highly intelligent conscious beings, but it is best to assume that not all highly intelligent conscious beings have such a conception. For these reasons, the ACT test should not be construed as a necessary condition that all AIs must pass. Put another way, failing ACT does not mean that a system is definitely not conscious. On the other hand, a system passing ACT should be regarded as conscious and be extended appropriate legal protections.

So, back to the superintelligent AI in the “box”—we watch and wait. Does it begin to philosophize about minds existing in addition to bodies, like Descartes? Does it dream, as in Isaac Asimov’s *Robot Dreams*? Does it express emotion, like Rachel in *Blade Runner*?Can it readily understand the human concepts that are grounded in our internal conscious experiences, such as those of the soul or atman? We suspect that the age of AI will be a time of soul-searching—both of ours, and for theirs.

Now let’s turn to a second test. Recall the thought experiment in which you had a neural rejuvenation at *Mindsculpt.* To introduce the chip test for synthetic consciousness, you are again the test subject. But it isn’t 2060; it is 2045, and the technology is still at the early stages of development. You have just learned that you have a brain tumor in your claustrum, one part of the brain responsible for the felt quality of your conscious experience. In a last ditch effort to survive, you enroll in a scientific study. You head to *iBrain,* hoping for the cure.

**The Chip Test**

**Recall that silicon-based brain chips are already under development as a treatment for various memory-related conditions, such as Alzheimer’s and PTSD, and both *Kernel* and *Neuralink* aim to develop AI-based brain enhancements for healthy individuals.**

In a similar vein, over at *iBrain*, researchers are trying to create chips that are functional isomorphs of parts of the brain, like the claustrum. They will gradually replace parts of your brain with brand new, durable microchips. As before, you are to stay awake during the surgery, and report any changes to the felt quality of your consciousness. The scientists are keen to learn whether any aspect of your consciousness is impaired. Their hope is to perfect neural prostheses in areas of the brain underlying consciousness.

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.

This would indicate a “substitution failure” of the artificial part for the original component. *Microchips of that sort just don’t seem to be the right stuff.*

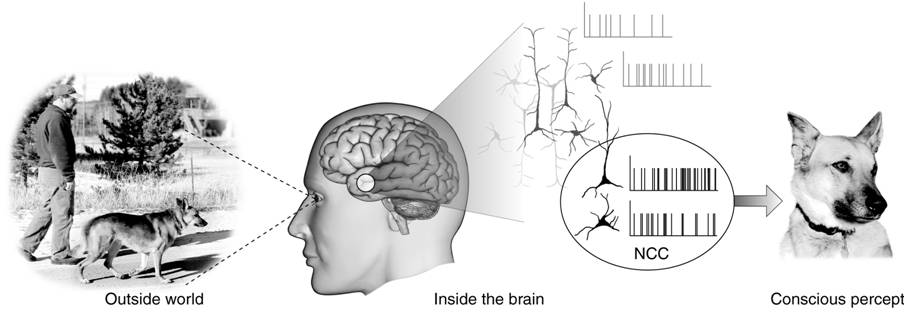
**This procedure would serve as a means of determining whether a chip made of a certain substrate and architecture can underwrite consciousness, at least when it is placed in a larger system that we already believe to be conscious.**

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 that science can eventually resolve? But after years and years of trying, we may reasonably question whether that kind of chip is a suitable substitute for carbon when it comes to consciousness.

Further, if science makes similar attempts with all other feasible substrates and architectures, a global failure would be a sign that for all intents and purposes, conscious AI isn’t possible. We may still regard conscious AI as conceivable, but from a practical standpoint – from the vantage point of our technological capacities – it is just isn’t possible. It may not even be compatible with the laws of nature to build consciousness into another, non-neural, substrate.

On the other hand, what if a certain kind of microchip works? In this case, we have reason to believe that this kind of chip is the “right stuff”, although it is important to bear in mind that our conclusion pertains to that specific microchip only. Further, even if a type of chip works in humans, there is still the further issue of whether the AI in question has the right functional organization for consciousness. We should not simply assume, even if chips work in humans, that all AIs that are built with these chips are conscious.

What is the value of a chip test then? It plays several important roles. First, it tells us when a substrate could serve as part of the basis of consciousness in a human. Depending upon where the neural prosthetic is placed, this may be a part of the brain responsible for a person’s ability to gate contents of consciousness, for one’s capacity for wakefulness or arousal (as with the brain stem), or it could be part or all of what is called the *neural correlate for consciousness*.[[25]](#endnote-25) A neural correlate for consciousness is the smallest set of neural structures or events that is sufficient for one’s having a memory or conscious percept, such as an image of the dog pictured below.



(Public domain. Have redrawn or cite from Wiki on “neural correlates of consciousness.” From Koch originally.)

Second, if a type of chip “passes” when it is embedded into a biological system, this alerts us to search carefully for consciousness in AIs that have these chips. Other tests for machine consciousness, such as the ACT test, can be administered, at least if the appropriate conditions for the use of such tests are met. If it turns out that only one kind of chip passes the chip test, and no other, it could be that being constructed of chips of this type is necessary for machine consciousness. (The requirement of this type of chip would be a “necessary condition” for synthetic consciousness, a requisite ingredient that conscious machines all have, like Hydrogen in H20.)

Both IIT and the chip test can suggest cases that ACT could miss. For instance, a nonlinguistic, highly sensory-based consciousness, like that of nonhuman animals, could be built from chips that pass the chip test. Or it could have a high Φ value, yet the AI may nevertheless lack the intellectual sophistication to pass an ACT test. It may even lack the behavioral markers of consciousness employed in a nonlinguistic version of ACT, such as mourning the dead.  But it could still be conscious.

Third, suppose a neurology patient’s conscious experience can be fully restored by a prosthetic chip placed in her hot zone.[[26]](#endnote-26) Such successes inform us about the level of functional connectivity that is needed for the neural basis of consciousness in that part of her brain. Further, it may help determine the level of functional detail that is needed to facilitate a sort of synthetic consciousness that is reverse engineered from the brain, although it may be that the "granularity" of the functional simulation may vary from one part of the brain to the next.

A well-known test for AI consciousness is the Integrated Information Theory (IIT), developed by the neuroscientist Guilio Tononi and his collaborators at the University of Wisconsin, Madison. There, they have been translating the felt quality of our experiences our into the language of mathematics.

**Integrated Information Theory (IIT)**

Tononi had an encounter with a patient in a vegetative state, and it convinced him that understanding consciousness merely a philosophical endeavor. “There are very practical things involved,” Tononi said to a *New York Times* reporter. “Are these patients feeling pain or not? You look at science, and basically science is telling you nothing.”[[27]](#endnote-27) Deeply intrigued by philosophy, his point of departure is the aforementioned hard problem of consciousness, which asks: how could matter give rise to the felt quality of experience?

Tononi’s answer is that consciousness requires a high level of “integrated information” within a system. Information is “integrated” within a system when its states are highly interdependent, featuring a rich web of feedback between its parts.[[28]](#endnote-28) The level of integrated information can be measured (such is designated by the Greek letter, Φ, pronounced “phi”). IIT holds that if we know the value of Φ, we can determine if a system is conscious and how conscious it is.

In the eyes of its proponents, IIT has the potential to serve as a test for synthetic consciousness: machines that have the requisite Φ level are conscious. Like the ACT test, IIT looks beyond superficial features of an AI, such as its humanlike appearance. Indeed, different kinds of AI architectures can be compared in terms of their measure of integrated information. The presence of a quantitative measure for phenomenal consciousness would be incredibly useful. Unfortunately, the calculations involved in computing Φ for even a small part of the brain, such as the claustrum, are computationally intractable. (That is, Φ can’t be calculated precisely except for extremely simple systems.)

Simpler metrics that approximate Φ have been provided, however, and the results are encouraging. For instance, the cerebellum has a relatively low Φ level, predicting that it contributes little to the overall consciousness of the brain. This fits with the data. Humans born without a cerebellum (a condition called “cerebellar agenesis”) do not seem to differ from normal subjects in the level and quality of their conscious lives. The cerebellum has low interconnectedness, exhibiting “feedforward” processing. In contrast, parts of the brain which, when injured or missing, contribute to a certain kind of loss in conscious experience have higher Φ values. IIT is also able to discriminate between levels of consciousness in normal humans (wakefulness versus sleep) and even single out “locked in” patents, who are unable to communicate.

IIT is what astrobiologists call a “small N” approach -- an approach that reasons from the biological case on Earth to a broader range of cases (in this case, the class of conscious machines). (“N” abbreviates “number”, so the name “small N” is referring to the small number of cases of a phenomenon at one’s disposal.) This is an understandable drawback, however, as the biological case is the only case of consciousness we know of. The tests I propose also have this drawback. Biological consciousness is the only case we know of, so we had better use it as our point of departure, together with a heavy dose of humility.

Another feature of IIT is that ascribes a small amount to consciousness to anything that has a minimal amount of Φ. In a sense, this is akin to the doctrine of panpsychism, a position on the nature of consciousness that we will discuss in the final chapter. According to this doctrine, microscopic and inanimate objects have at least a small amount of experience. But this does not mean that the view is panpsychist, at least if panpsychism is construed as claiming that everything has at least a small amount of experience. For IIT does not ascribe consciousness to everything. In fact, IIT does not predict that feedforward computational networks are conscious. As Tononi and Koch note: “ITT predicts that consciousness is graded, is common among biological organisms and can occur in some very simple systems. Conversely, it predicts that feed-for- ward networks, even complex ones, are not conscious, nor are aggregates such as groups of individuals or heaps of sand.” (Tononi and Koch, 2018)

IIT singles out certain systems as conscious in a special sense, however. That is, it aims to predict which systems have a more complex form of consciousness, akin to that what occurs in normally functioning brains.[[29]](#endnote-29) The question of AI consciousness, in this context, seeks to determine whether machines have macroconsciousness, as opposed to smaller Φ levels exhibited by everyday objects.

Is having high Φ sufficient for a machine’s being conscious? According to Scott Aaronson, the Director of the Quantum Information Center at the University of Texas at Austin, a two-dimensional grid that runs error-correction codes such as those used for CDs will have a very high Φ level. Aaronson writes: “… IIT predicts not merely that these systems are “slightly” conscious (which would be fine) but that they can be unboundedly more conscious than humans are.” But a grid does not seem to be the sort of thing that is conscious, suggesting to many that IIT should not be regarded as being sufficient for consciousness.

Tononi has responded to Aaronson’s point, biting the bullet, responding that the grid is conscious (i.e., macroconscious. I prefer instead to reject the view that having a high Φ value is *sufficient* for an AI to be conscious. But perhaps IIT could supply a necessary feature that all conscious systems have. One who finds IIT on the right track will likely suspect that for all we know, all conscious systems, biological or mechanical, may have the requisite minimum level of phi.

I find this dangerous. IIT is still highly speculative. For instance, consider that even today’s fastest supercomputers have low phi. This is because their chip designs are currently insufficiently neuromorphic. (Even machines using IBM’s True North chip, which is designed to be neuromorphic, have low phi, because True North has a bus.) It could very well be the case that systems running on computer hardware that has low phi, like the Summit supercomputer, but which eventually implement highly sophisticated informational structure, reverse engineering parts of the brain, implements a conscious system, perhaps akin to that of the level of consciousness of a mouse.

So until we know more, how are we to deal with a machine that has high Φ, should we ever encounter one?[[30]](#endnote-30) We’ve seen that Φ is probably not sufficient. Further, since research on Φ has been in biological systems and today’s computers that are not good candidates for being conscious, it is too early to tell whether Φ is a necessary condition for AI consciousness. It may still be a marker for consciousness—a feature that indicates that we should treat the system with special care as potentially being a conscious system.

There is a more general issue here that we need to deal with. The tests we’ve discussed are still under development, so there may be AIs that we suspect are conscious, but we are uncertain whether they are. To add to this uncertainty, I’ve stressed that the social impact of synthetic consciousness depends upon a number of variables. For instance, consciousness in one kind of machine may lead to empathy, but it may lead to volatility in a different machine architecture. So how should we proceed when IIT and the chip test identifies a marker for synthetic consciousness, or when ACT says an AI might be conscious? Should we proceed with the development of artificial consciousness? It depends. Here, I’ll suggest a precautionary approach.

**The Precautionary Principle**

Throughout this book, I’ve stressed that using several different indicators for AI consciousness is prudent; in the right circumstances, one or more tests can be used to check the results of another test, indicating deficiencies and avenues for improvement in testing.[[31]](#endnote-31) Perhaps, for instance, the microchips that pass the chip test are not those that IIT says have a high Φ value, or suppose that those chips that IIT predicts will support consciousness actually fail when used as neural prosthetics in the human brain.

The Precautionary Principle is a familiar ethical principle. It says that if there’s a chance of a technology causing catastrophic harm, it is far better to be safe than sorry. Before using a technology that could have a catastrophic impact on society, those wanting to develop that technology must first prove that it will not have this dire impact. Precautionary thinking has a long history, although the principle itself is relatively new. *The Late Lessons from Early Warnings Report* gives an example of a physician who recommended removing the handle of a water pump in London to stop a cholera epidemic in 1854. Although the evidence for the causal link between the pump and the spread of cholera was weak, the simple measure effectively halted the spread of cholera (Harremoes et al, 2001). Heeding the early warnings of the potential harms of asbestos would have saved many lives, although the science at this time was uncertain. According to a UNESCO report (2005), the precautionary principle has been a rationale for a large number of treaties and declarations in environmental protection, sustainable development, food safety and health.

I’ve emphasized the possible ethical implications of synthetic consciousness, stressing that at this time, we do not know whether conscious machines will be created, and what its impact on society would be. This means that developing tests for machine consciousness and gauging the impact of consciousness on other key features of the machine, such as empathy and trustworthiness, are needed. A precautionary stance suggests that we shouldn’t simply press on with the development of sophisticated AI without carefully gauging its consciousness and determining that it is safe. For the inadvertent or intentional development of conscious machines could carry so called existential risks to humans, risks ranging from volatile superintelligences that supplant humans to a human merger with AI, that diminishes or ends human consciousness.

In light of this, I offer three recommendations. First, ongoing testing, as indicated above. Second, if AI consciousness is found, and it compromises AI safety, it simply shouldn’t be developed in the architectural environments in which it was found to be unsafe. Third, if there is any doubt about an AI’s consciousness or safety, it shouldn’t be deployed in situations where it has the potential to cause catastrophic harm. Fourth, we are uncertain whether a given type of AI is conscious, but we have some reason to believe it may be, even in absence of a definitive test, a precautionary stance says we should extend the same legal protections to the AI that we extend to other sentient beings. For all we know, AI will have the capacity to suffer, like nonhuman animals do.Excluding conscious AIs from ethical consideration is speciesist, to borrow an expression that Peter Singer used in the context of the animal liberation movement. And it is better to be safe than sorry.

For instance, AIs made of chips that pass the chip test, even if they do not pass an ACT, should be regarded as having a “marker” for consciousness—a feature suggestive of consciousness. I’ve also mentioned that functional consciousness may be a marker for phenomenal consciousness. Projects working with AIs that have both of these markers could, for all we know, involve conscious AIs. Until we know whether these systems are conscious, it is best to treat them as if they are.

**Further Terrain: Exploring the Mind-Machine Merger**

So we’ve seen that while successful AI-based technologies obviously require solid scientific footing, their proper use involves rich philosophical issues that call for multidisciplinary collaboration, careful testing, and public dialogue. These matters cannot be settled by science alone. As the book unfolds, subsequent chapters shall illustrate other ways in which heeding this general observation may be key to the future of mind.

Recall the Jetsons Fallacy. **AI will not just make for better robots and supercomputers.** In reality, AI will not just transform the world around us, AI will transform us. The artificial hippocampus, brain chips to treat mental illness, neural lace -- these are just some of the transformative technologies under development today. Remember, *The Center for Mind Design* is not that far-fetched. So in the next several chapters we shall turn our gaze inward, exploring the idea of merging with AI. As we shall see, this is no simple decision.

# For instance, now that we’ve explored machine consciousness, we can appreciate that we really do not know whether AI components could effectively replace parts of the brain responsible for consciousness. In this domain, AI research may hit a wall. If so, then humans cannot fully merge with AI in a safe manner, for they would lose that which is most central to their lives – their consciousness. In this case, AI-based enhancements would have to be limited to parts of the brain that are not responsible for consciousness. Only biological enhancements could be used in areas of the brain responsible for consciousness.

# Alternately, perhaps nanoscale AI-based enhancements in these areas is still possible, if they merely complement the processing of these brain areas, while not replacing neural tissue or interfering with conscious experience. Notice that in any of these cases, a merger or fusion with AI is not in the cards, although a limited integration is still possible. In either scenario, one couldn’t upload to the cloud or replace all of one’s neural tissue with AI components.

# These are emerging technologies, so we cannot really tell how things will unfold. But for the sake of discussion, let’s suppose that AI-based enhancements *can* replace parts of the brain responsible for consciousness. Even so, as the next chapters illustrate, there are reasons to resist merging with AI. For all we know, slowly enhancing through cumulatively significant alterations to the brain and/or by moving from one substrate to another without even augmenting one’s intelligence, will *not* clearly preserve you. As before, we will begin with a fictional scenario, one designed to help you consider the virtues and vices of radical enhancement.

**Chapter Five: Could You Merge with AI?**

*Suppose it is 2035 and being a technophile, you decide to add a mobile internet connection to your retina.  A year later, you enhance your working memory by adding neural circuitry. You are now officially a cyborg. Now skip ahead to 2045. 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.” Posthumans are future beings who are no longer unambiguously human, for they have mental capacities that radically exceed those of present day humans. 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.*

*It is now 2300 AD. Worldwide technological developments, including your own enhancements, are facilitated by superintelligent AI. Recall that a superintelligent AI has the capacity to radically outperform the best human brains in practically every field, including scientific creativity, general wisdom, and social skills. Over time, the slow addition of better and better AI components has left no real intellectual difference in kind between you and a superintelligent AI. The only 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. You have merged with AI.*

The thought experiment features the kind of enhancements that transhumanists and certain well-known tech leaders, such as Elon Musk and Ray Kurzweil, aspire to.[[32]](#endnote-32) Recall that transhumanists aim to redesign the human condition, striving for immortality and synthetic intelligence, all in hopes of improving our overall quality of life. Proponents of the idea that humans should merge with AI are techno-optimists, for they hold that synthetic consciousness is possible. In addition, they believe a merger or fusion with AI is possible. More specifically, they tend to suggest the following trajectory of enhancements: [[33]](#endnote-33)

21st century unenhanced human → significant “upgrading” with cognitive and other physical enhancements→ posthuman status → “superintelligent AI”[[34]](#endnote-34)

Let us call the view that humans should follow such a trajectory, and merge with AI, “fusion optimism.” Techno-optimism about machine consciousness doesn’t require a belief in fusion-optimism, although many techno-optimists are sympathetic to the view. But fusion optimism aims for a future in which these posthumans are conscious beings.

So, should you embark upon this journey? [[35]](#endnote-35) Unfortunately, as attractive as superhuman abilities may sound, we’ll see that radical, and even mild brain enhancements could turn out to be risky, not resulting in the preservation of one’s original self. Each human who embarks upon these enhancements, unbeknownst to them, could end their life in the process. The being that is created by the “enhancement” procedure is someone else entirely. This is hardly an enhancement at all. The so called “enhancement” would result in the death of every human who tries to merge with AI.

**What is a Person?**

For consider that in order to understand whether you should enhance yourself, you must first understand what you are to begin with. But what is a person? And, given your conception of a person, after such radical changes, would you yourself continue to exist, or would you have ceased to exist, having been replaced by someone else? If the latter is the case, why would you want to attempt to merge with AI?

To make such a decision, you must understand the metaphysics of personal identity – that is, you must answer the question: What is it in virtue of which a particular self or person continues existing over time? One way to begin appreciating the issue is to consider the persistence of everyday objects. Consider the espresso machine in your favorite café. Suppose that five minutes have elapsed and the barista turns the machine off. Imagine asking her if the machine is the same one that was there five minutes ago. She will likely tell you the answer is glaringly obvious. It is of course possible for one and the same machine to continue existing overtime. This seems to be a reasonable case of the persistence of the machine, even though at least one of the machine’s features (or “properties”) has changed. On the other hand, if the machine disintegrates or melts, then the same machine would no longer exist. What remained wouldn’t be an espresso machine at all, for that matter.

The point is that when it comes to the objects around us, some changes cause a thing to cease to exist, while others do not. Philosophers call the characteristics that a thing must have as long as it exists “essential properties.” (By “property” philosophers mean features, such as the flatness of the table, or the blueness of the sky.)

Now let’s reconsider the transhumanist’s trajectory for enhancement: for radical enhancement to be a worthwhile option, it should be a form of personal development. At bare minimum, even if enhancement brings such goodies as superhuman intelligence and radical life extension, it must not involve the elimination of any of your essential properties. For if that happened, the sharper mind and fitter body would not be experienced by you – they would be experienced by someone else.

Even if you would like to be superintelligent, knowingly embarking on a path that trades away one or more of your essential properties would be tantamount to suicide – that is, to your intentionally causing yourself to cease to exist. So before you enhance, you had better get a handle on what your essential properties are.

Let’s mull over what your essential properties might be. Think of yourself in first grade. What properties have persisted that seem somehow important to you still being the same person? Notice that the cells in your body have now changed, and your brain structure and function has altered dramatically. If you are simply the physical stuff that comprised your brain and body in first grade, you would have ceased to exist some time ago. That physical first-grader is simply not here, any longer. Kurzweil clearly appreciates the difficulties here, commenting:

So who am I? Since I am constantly changing, am I just a pattern? What if someone copies that pattern? Am I the original and/or the copy? Perhaps I am this stuff here – that is, the both ordered and chaotic collection of molecules that make up my body and brain. [[36]](#endnote-36)

Kurzweil is referring to two theories that are center stage in the age‐old philosophical debate over the nature of persons. The leading theories include the following:

1. The soul theory: your essence is your soul or mind, understood as a nonphysical entity distinct from your body.
2. The psychological continuity theory: you are essentially your memories and ability to reflect on yourself (Locke) and, in its most general form, you are your overall psychological configuration, what Kurzweil referred to as your “pattern.”[[37]](#endnote-37)
3. Brain-based Materialism: you are essentially the material that you are made out of, i.e., your body and brain – what Kurzweil referred to as “the ordered and chaotic collection of molecules” that make up my body and brain.[[38]](#endnote-38)[[39]](#endnote-39)
4. The no self view: the self is an illusion. The “I” is a grammatical fiction (Nietzsche). There are bundles of impressions but there is no underlying self (Hume). There is no survival because there is no person (Buddha).[[40]](#endnote-40)

Each of these views has its own implications about whether to enhance. For instance, suppose you hold the first view, a position held by adherents to Christianity, for instance, as well as by the famous philosopher Rene Descartes[[41]](#endnote-41). In this case, your decision to enhance would seem to depend on whether you have justification for believing that the enhanced body would retain your soul or immaterial mind.

Now suppose that you are a proponent of brain-based materialism, the third view. Views that are materialist hold that minds are basically physical or material in nature and that mental features, such as the thought that espresso has a wonderful aroma, are ultimately just physical features. (This view is often called “physicalism” as well.) Brain-based materialism says this, and in addition, it ventures the further claim that your thinking is dependent upon the brain—thought cannot “transfer” to a different substrate. So on this view, enhancements must not change one’s material substrate, or the person would cease to exist.

The third view, in contrast, holds that enhancements can alter the kind of substrate but they must preserve your overall psychological configuration. This view, unlike brain-based materialism, would allow you to transition to silicon or some other substrate, at least in principle.

Finally, the fourth position contrasts sharply with the others. If you hold the No Self View, then the survival of the person is not an issue, for there is no person or self there to begin with. In this case, expressions like “I” and “you” do not really refer to persons or selves, and Nietzsche was right. Notice that if you are a proponent of the No Self View, you may strive to enhance nonetheless. For instance, you might find intrinsic value in adding more superintelligence to the universe – you might value life forms with higher forms of consciousness and wish that your “successor” be such a creature.

I don’t know whether many of those who publicize the idea of a mind-machine merger, such as Elon Musk and Michio Kaku, have considered these classic positions on personal identity. But they should. It is a bad idea to ignore this debate. One could be dismayed, at some later point, to learn that a technology one advocated actually had a tremendously negative impact on human flourishing.

In any case, both Kurzweil and Bostrom have considered the issue in their work. They, like many other transhumanists, adopt a novel and intriguing version of the psychological continuity view; in particular, they adopt a computational, or *patternist*, account of continuity.

**Are you a Software Pattern**?

Patternism’s point of departure is empirical work in cognitive science that suggests the brain is computational. Research in cognitive science is commonly regarded as motivating a view called “the computational theory of mind” (CTM), a general position that holds thinking can be explained as computation. Originally, versions of CTM held that the mind is akin to a standard computer, but nowadays, it is commonly agreed that the brain does not have the structure of a standard computer. But cognitive and perceptual capacities, such as working memory and attention, are still considered computational. Although computational theories of mind differ in the details, one thing they have in common is that they all explain cognitive and perceptual capacities in terms of causal relationships between components, each of which can be described algorithmically. One common way of describing CTM is by reference to the idea that the mind is a software program. That is:

The Software Approach to the Mind (SAM). The mind is the program running on the hardware of the brain. That is, the mind is the algorithm the brain implements, where this algorithm is something that cognitive science discovers, at least in principle.[[42]](#endnote-42)

Those working on computational theories of mind in philosophy of mind tend to ignore the topic of patternism, as well as the more general topic of personal identity. This is unfortunate for two reasons. First, on any feasible view of the nature of persons, one’s view of the nature of mind plays an important role. For what is a person if not, at least in part, that which she thinks and reflects with? Second, whatever the mind is, an understanding of its nature should include the study of its persistence, and this sort of undertaking would be closely related to theories of the persistence of the self or person. Yet the issue of persistence is often ignored. I suspect the reason is simply that work on the nature of the mind is in a different subfield of philosophy from work on the nature of the person.

Transhumanists step up to the plate in trying to connect the topic of the nature of the mind with issues regarding personal identity, and they are clearly right to sense an affinity between patternism and SAM. After all, if you take a computational approach to the nature of mind, it is natural to regard persons as being somehow computational in nature, and to ponder whether the survival of a person is somehow a matter of the survival of their software pattern. The guiding conception of the patternist is aptly captured by Ray Kurzweil:

The specific set of particles that my body and brain comprise are in fact completely different from the atoms and molecules that I comprised only a short while ago. We know that most of our cells are turned over in a matter of weeks, and even our neurons, which persist as distinct cells for a relatively long time, nonetheless change all of their constituent molecules within a month… . I am rather like the pattern that water makes in a stream as it rushes past the rocks in its path. The actual molecules of water change every millisecond, but the pattern persists for hours or even years. [[43]](#endnote-43)

Later in his discussion, Kurzweil calls his view “Patternism” (ibid.: 386). Put in the language of cognitive science, as the transhumanist surely would, what is essential to you is your computational configuration – for example, what sensory systems/subsystems your brain has (e.g. early vision), the way that the basic sensory subsystems are integrated in the association areas, the neural circuitry making up your domain general reasoning, your attentional system, your memories, and so on – overall, the algorithm that your brain computes.12

You might think the transhumanist views a brain-based materialism favorably. Transhumanists generally reject brain-based materialism, however, for they tend to believe the same person can continue to exist if her pattern persists, even if she is an upload, no longer having a brain.

Of course, I’m not suggesting all transhumanists are patternists.[[44]](#endnote-44) But Kurzweil’s patternism is highly typical of transhumanism. For instance, consider the appeal to patternism in the following passage of *The Transhumanist Frequently Asked Questions*, of which Nick Bostrom is the primary author. It begins by discussing the process of uploading:

Uploading (sometimes called “downloading,” “mind uploading” or “brain reconstruction”) is the process of transferring an intellect from a biological brain to a computer. One way of doing this might be by first scanning the synaptic structure of a particular brain and then implementing the same computations in an electronic medium. … An upload could have a virtual (simulated) body giving the same sensations and the same possibilities for interaction as a non‐simulated body. … Advantages of being an upload would include: Uploads would not be subject to biological senescence. Backup copies of uploads could be created regularly so that you could be rebooted if something bad happened. (Thus your lifespan would potentially be as long as the universe’s.) … Radical cognitive enhancements would likely be easier to implement in an upload than in an organic brain. … A widely accepted position is that you survive so long as certain information patterns are conserved, such as your memories, values, attitudes, and emotional dispositions … For the continuation of personhood, on this view, it matters little whether you are implemented on a silicon chip inside a computer or in that gray, cheesy lump inside your skull, assuming both implementations are conscious. (Bostrom 2003)

In short, the transhumanist’s futuristic, computationalist orientation leads them to *patternism*: an approach to the nature of persons that is an intriguing blend between the computational approach to the mind and the traditional psychological continuity view of personhood.[[45]](#endnote-45) If plausible, patternism would explain how one can survive such radical enhancements as those depicted in our thought experiments. Further, it would be an important contribution to the age‐old philosophical debate over the nature of persons. So, is it correct? And further, is patternism even compatible with the radical enhancements fusion optimists envision? In the next chapter, we shall consider these issues.

# Chapter 6 Getting a Mindscan

# I teach you the Overman! Mankind is something to be overcome. What have you done to overcome mankind?

# -Fredrick Neitzsche, *Thus Spoke Zarathusthra*

Fusion optimists claim that radical enhancements are compatible with survival. You and I, we are just informational patterns, and we can be upgraded to a new, superior version -- human 2.0, if you will. And from there, as AI developments continue, further versions of us can be created, until one day, when the science is just right, in the ultimate Nietzschean act of self-overcoming, we merge with AI.

I’ve suggested that radical, and even mild, brain enhancements could be risky, not resulting in the preservation of one’s original self. Each human who embarks upon these enhancements, unbeknownst to them, could end their life in the process. In that case, the being that is created by the “enhancement” procedure is someone else entirely.

Let’s consider whether the fusion-optimists are right by thinking about a scenario depicted in science fiction novel called *Mindscan*, by Robert Sawyer. In Sawyer’s novel, the protagonist Jake Sullivan has an inoperable brain tumor. Death could strike him at any moment. Luckily, Immortex has a new cure for aging and illness – a “mindscan.” Immortex scientists will upload his brain configuration into a computer and “transfer” it into an android body that is designed using his own body as a template. Although imperfect, the android body has its advantages – once an individual is uploaded, a backup exists that can be downloaded if one has an accident. And it can be upgraded as new developments emerge. Jake will be immortal.

Sullivan enthusiastically signs numerous legal agreements. He is told that, upon uploading, his possessions will be transferred to the android, who will be the new bearer of his consciousness. Sullivan’s original copy, which will die soon anyway, will live out the remainder of his life on “High Eden,” an Immortex colony on the moon. Although stripped of his legal identity, the original copy will be comfortable there, socializing with the other originals who are also still confined to biological senescence.

Sawyer then depicts Jake’s experience while lying in the scanning tube a few seconds before the scan. Sawyer writes (from Jake’s perspective):

“I was looking forward to my new existence. Quantity of life didn’t matter that much to me – but quality. And to have time – not only years spreading out into the future, but time in each day. Uploads, after all, didn’t have to sleep, so not only did we get all those extra years, we got one‐third more productive time. The future was at hand. Creating another me. Mindscan.

But then, a few seconds later:

‘All right, Mr. Sullivan, you can come out now.’ It was Dr. Killian’s voice, with its Jamaican lilt.

My heart sank. No …

‘Mr. Sullivan? We’ve finished the scanning. If you’ll press the red button …’ It hit me like a ton of bricks, like a tidal wave of blood. No! I should be somewhere else, but I wasn’t… .

I reflexively brought up my hands, patting my chest, feeling the softness of it, feeling it raise and fall. Jesus Christ!

I shook my head. ‘You just scanned my consciousness, making a duplicate of my mind, right?’ My voice was sneering. “And since I’m aware of things after you finished the scanning, that means I – this version – isn’t that copy. The copy doesn’t have to worry about becoming a vegetable anymore. It’s free. Finally and at last, it’s free of everything that’s been hanging over my head for the last twenty‐seven years. We’ve diverged now, and the cured me has started down its path. But this me is still doomed… .” [[46]](#endnote-46)”

Sawyer’s novel is a *reductio ad absurdum* of the patternist conception of the person. For all that patternism says is that as long as person A has the same computational configuration as person B, A and B are the same person. Indeed, Sugiyama, the person selling the mindscan to Jake, had espoused a form of patternism.[[47]](#endnote-47)

Jake’s unfortunate experience can be put into the form of a challenge to patternism, which we shall call “the reduplication problem”: only one person can really be Jake Sullivan, as Sullivan reluctantly found out. But according to patternism, both creatures are Jake Sullivan – for they share the very same psychological configuration. But, as Jake learned, while the creature created by the mindscan process may be a person, it is not the *very same* person as Jake. It is just another person with an artificial brain and body configured like the original. Hence, having a particular type of pattern cannot be *sufficient* for personal identity. Indeed, the problem is illustrated to epic proportions later in the book when numerous copies of Sullivan are made, all believing they are the original! Ethical and legal problems abound.

# A Way Out?

The patternist has a response to all this, however. As noted, the reduplication problem suggests that sameness of pattern is not sufficient for sameness of person. You are more than just your pattern. But there still seems to be something right about patternism – for as Kurzweil notes, throughout the course of your life your cells change continually; it is your organizational pattern that carries on. Unless you have a religious conception of the person, and adopt the soul theory, patternism may strike you as inevitable, at least insofar as you believe there is such a thing as a person to begin with.

In light of this, perhaps we should react to the reduplication case in the following way: your pattern is *essential* to yourself despite not being *sufficient for* a complete account of one’s identity. Perhaps there is an additional essential property which, together with your pattern, yields a complete theory of personal identity.

What could the missing ingredient be? Intuitively, it must be a requirement that serves to rule out mindscans and, more generally, any cases in which the mind is “uploaded.” For any sort of uploading case will give rise to a reduplication problem, for uploaded minds can in principle be downloaded again and again.

Now think about your own existence in space and time. When you go out to get the mail, you move from one spatial location to the next, tracing a path in space. A spacetime diagram can help us visualize the path one takes throughout one’s life. Collapsing the three spatial dimensions into one (the vertical axis) and taking the horizontal axis to signify time, consider the following typical trajectory:

**[Insert figure of a worm in spacetime]**

Notice that the figure carved out looks like a worm; you, like all physical objects, carve out a sort of “spacetime worm” over the course of your existence.

This, at least, is the kind of path that “normals” – those who are neither posthumans nor superintelligences – carve out. But now consider what happened during the mindscan. Again, according to patternism, there would be two of the very same person. The copy’s spacetime diagram would look like the following:

**[Figure of a different worm]**

This is bizarre. It appears that Jake Sullivan exists for 42 years, has a scan, and then somehow instantaneously moves to a different location in space and lives out the rest of his life. This is radically unlike normal survival. This alerts us that something is wrong with pure patternism: it lacks a requirement for spatiotemporal continuity.

This additional requirement would seem to solve the reduplication problem. For consider the day of the mindscan. Jake went into the laboratory and had a scan; then he left the laboratory and went directly into a spaceship and flew to Mars. It is this man – the one who traces a continuous trajectory through space and time – who is the true Jake Sullivan.

This response to the reduplication problem only goes so far, however. For consider Sugiyama, who, when selling his mindscan product, ventured a patternist pitch. If Sugiyama had espoused patternism together with a spatiotemporal continuity clause, few would have signed up for the scan. For that extra ingredient would rule out a mindscan, or any kind of uploading for that matter, as a form of survival. Only those wishing to have a mere replacement for themselves would sign up.

There is a general lesson here for the transhumanist or fusion-optimist: if one opts for patternism, enhancements like uploading to avoid death or to facilitate further enhancements are not really “enhancements” but forms of suicide. *The fusion-optimist should sober up and not offer such procedures as enhancements*. When it comes to enhancement, there are intrinsic limits to what technology can deliver. (Ironically, the proponent of the soul theory is in better shape here. For perhaps the soul does teleport. Who knows?)

Now let’s pause and take a breath. We’ve accomplished a lot in this chapter. We began by thinking about the Mindscan case, and a “reduplication problem” arose for patternism. This led me to discard the original form of patternism as false. I then suggested a way to modify patternism, to arrive at a more realistic position. This meant adding a new element to the view, the “spatiotemporal continuity clause”, which required that there be spatiotemporal continuity in a pattern for the original pattern to survive. We called this *modified patternism*. Incidentally, in the personal identity debates in metaphysics, appealing to spatiotemporal continuity is a common reaction to problems arising for the related psychological continuity view.

Modified patternism may strike you more sensible, but notice that it doesn’t serve the fusion-optimist well, for according to modified patternism, uploading is incompatible with survival. Remember, uploading violates the spatiotemporal continuity requirement on persistence – a requirement that modified patternism adopted. But what about other, fairly radical, AI-based enhancements? Are these ruled out as well? Consider, for instance, selecting a bundle of enhancements at the Center for Mind Design. These enhancements could dramatically alter your mental life, yet they do not involve uploading, and it is not obvious that spatiotemporal continuity would be violated.

Indeed, the fusion-optimist could point out that one could still merge with AI through a series of gradual, but cumulatively significant enhancements that added AI-based components inside the head, slowly replacing neural tissue. This wouldn’t be “uploading” because one’s thinking would still be inside the head, but the series still amounts to an attempt to transfer one’s mental life to another substrate. For when the series was completed, if it worked, the individual’s mental life would have migrated from a biological substrate to a nonbiological one, such as silicon. And the fusion optimist would be right: humans can merge with AI.

But would it work? Here, we need reconsider some issues raised in the previous chapter.

# Death or Personal Growth?

Suppose you are at the Center for Mind Design, and you are gazing at the menu, considering buying a certain bundle of enhancements. But now, suppose you’ve read this book. Longing to upgrade yourself, you reluctantly consider whether modified patternism might be true. And you wonder: If I am a modified pattern, what happens to me when I add the enhancement bundle? My pattern will change, so do I die?

To determine if this would be the case, the modified patternist would need to give a more precise account of what a “pattern” is, and when different enhancements do and do not constitute a continuation in the pattern. The extreme cases seem clear – for instance, as discussed, mindscans are ruled out by the spatiotemporal continuity requirement. And further, because both versions of patternism are versions of the older psychological continuity approach, the modified patternist will likely want to say that a memory erasure process that erased a difficult year from one’s childhood is an unacceptable alteration of one’s pattern, removing too many of one’s memories and changing the person’s nature. On the other hand, mere everyday cellular maintenance by nanobots to overcome the slow effects of aging would, presumably, not affect the identity of the person, for it wouldn’t alter one’s memories.

The problem is that the middle range cases are unclear. Maybe deleting a few bad chess‐playing habits is kosher, but what about more serious mindsculpting endeavors, like the enhancement bundle you were considering, or even adding a single cognitive capacity, for that matter? Or what about raising your IQ by 30 points, or erasing all of your memory of some personal relationship, as in the film Eternal Sunshine of the Spotless Mind? The path to superintelligence may very well be at the end of a gradual path through a series of these smaller sorts of enhancements. But where do we draw the line?

Each of these enhancements are less radical than uploading -- perhaps we could regard them as “middle range” enhancements -- but each one could be an alteration in the pattern that is incompatible with the preservation of the original person. And cumulatively, their impact on one’s pattern and can be quite significant. So again, what is needed is a clear conception of what a pattern is, and what changes in pattern are acceptable and why. Without a firm handle on this issue, the transhumanist developmental trajectory is, for all we know, the technophile’s alluring path to suicide.

This problem looks hard to solve in a way that is compatible with preserving the very idea that we can be identical over time to some previous or future self. Determining a boundary point threatens to be an arbitrary exercise in which once a boundary is selected, an example is provided suggesting the boundary should be pushed outward, *ad nauseam*. On the other hand, there is something insightful about the view that over time one gradually becomes less and less like one’s earlier self. But appreciate this point too long and it may lead to a dark place: for if one finds patternism or modified patternism compelling to begin with, how is it that one truly persists over time, from the point of infancy until maturity, during which time there are often major changes in one’s memories, personality, and so on?

Indeed, even a series of gradual changes cumulatively amounts to an individual, B, who is greatly altered from her childhood self, A. Why is there really a relation of identity that holds between A and B, instead of an ancestral relation: A’s being the ancestor of B? Put differently, how can we tell if that future being who exists, after all these enhancements is really us, and not, instead, a different person—a sort of “descendent” of ours?

It is worth pausing for a moment to reflect on the relationship between the ancestor and his or her “descendant,” although it a bit of an aside. There is clearly an intriguing and intimate connection here. The connection resembles a parent-child relationship, but in some ways, it is more intimate, for you have first-person knowledge of the past of this new being. You, and not this person, have literally lived those past moments. In contrast, we may feel closely connected to our children’s lives, but we do not literally see the world through their eyes. But in another sense, the relationship is weaker than many parent-child connections. For you two will never even be in the same room. Like a woman who dies in childbirth, you will never meet your “descendant.”

Perhaps, your mindchild will come to mourn your passing, regarding you with deep affection and appreciating that the end of your live was responsible for the beginning of their own. But that assumes that understand that they are not really you. They may live for millennia, yet they may never come to appreciate that you are distinct from them, and that you ceased to exist. This is a unique and complex relationship indeed.

To the extent that the enhanced being is still much like you, you may, in anticipation of their creation, feel a special connection to his or her interests, memories, and so on. You may even feel a special connection to a being who will be nothing like you. For instance, in an act of benevolent mindsculpting, perhaps you would deliberately set out to create a superintelligent being, knowing that if you succeed, you will die.

In any case, the main point of this section was to show you that even modified patternism faces a key challenge—it needs to tell us when a shift in one’s pattern is compatible with survival, and when it is not. I’ve also mentioned that there is a second problem with modified patternism as well. This second issue also relates to the issue of gradual, but cumulatively significant, changes in one’s pattern.

**Leaving your Substrate for Another?**

I’ve mentioned that a second problem arises for modified pattern as well, one that challenges the very possibility that one could transfer to a different substrate, even with no cognitive or perceptual enhancements.

Suppose that it is 2050, and people are getting gradual neural regeneration procedures as they sleep. During their nightly slumbers, nanobots slowly import nanoscale materials that are computationally identical to the original materials. The nanobots then gradually remove the old materials, setting them in a small container beside the person’s bed.

By itself, this process is unproblematic for modified patternism. But now suppose there is an optional upgrade to the regeneration service for those who would like to make a backup copy of their brains. If one opts for this procedure, then, during the nightly process, the nanobots take the replaced materials out of the dish and place them inside a cryogenically frozen biological brain.

[add illustration here]

Suppose, that by the end of the process the materials in the frozen brain have been entirely replaced by the person’s original neurons, and they are configured the same way that they were in the original brain.

Now, suppose you choose to undergo this procedure alongside your nightly regeneration. Over time, this second brain comes to be composed of the very same material as your brain originally was, configured in precisely the same manner. Which one is you? The original brain, which now has entirely different neurons, or the one with all your original neurons? The modified patternist has this to say about the neural regeneration case: you are the creature with the brain with entirely different matter, as this creature traces a continuous path through spacetime. But now, things go awry: why is spatiotemporal continuity supposed to outweigh other factors, like being composed of the original material substrate? Here, to be blunt, my intuitions crap out. I do not know if this thought experiment depicts a scenario that is technologically feasible or that is even compatible with the laws of nature, but nevertheless, it raises a deep point. We are trying to find out what the essential nature of a person is, so we’d like to find a solid justification for selecting one option above the other. Assuming there is a persisting self, which is the deciding factor that enables one to survive, in case in which psychological continuity holds—being made of one’s original parts or spatiotemporal continuity?

These problems suggest that modified patternism needs a good deal of spelling out. And remember, it wasn’t consistent with uploading in any case. The original patternist view held that one can survive uploading, but we discarded the view as deeply problematic. Until the fusion optimist provides a solid justification for his position, it is best to view the idea of merging with AI with a good deal of skepticism. Indeed, we should even question forms of brain enhancement that stop short of an attempt to merge with AI. Enhancements that merely involve the rapid or even gradual replacement of parts of one’s brain may be risky.

**Metaphysical Humility**

At the outset of the book, I asked you to imagine a shopping trip at *The Center for Mind Design*. You can now see how deceptively simple this little thought experiment was. Perhaps the best response to the ongoing controversy over the nature of persons is to take a simple stance of *metaphysical humility*. Claims about survival that involve a person “transferring” one’s mind to a new type of substrate or a person making drastic alterations to one’s brain should be scrutinized. Answers to whether such ‘enhancements’ are truly compatible with survival cannot be found by consulting science alone. As alluring as superintelligence or digital immortality may be, we’ve seen that there is much disagreement in the personal identity literature over whether any of these ‘enhancements’ are a means of survival. One is, of course, free to disregard my warnings and be a risk taker. But remember, enhancements are optional; a person can lose her life by making a bad enhancement decision.

A stance of metaphysical humility says the way forward is public dialogue, informed by metaphysical theorizing. At its best, a pluralistic society should recognize the diversity of different views on these matters, and not assume science itself can answer questions about whether radical forms of brain enhancement are a form of survival.

Nevertheless, as *The Transhumanist Frequently Asked Questions* indicates, the development of enhancements like brain uploading or adding brain chips to augment intelligence or radically alter one’s perceptual abilities are key enhancements invoked by the transhumanist view of the development of the person. Such enhancements sound strangely like the thought experiments philosophers have used for years as problem cases for various theories of the nature of persons, so it doesn’t surprise me a bit that deep problems emerge.

We’ve learned that the Mindscan example suggests that one should not upload, and that the patternist needs to modify her theory to rule out such cases. Even with this modification in hand, however, transhumanism and fusion optimism still require a detailed account of what constitutes a break in a pattern versus a mere continuation of it. Without progress on this issue, it will not be clear if medium range enhancements, such adding neural circuitry to make oneself smarter, are safe. Finally, the nanobot case warns against migrating to a different substrate, even if one’s mental abilities remain unchanged. Given all this, it is fair to say that the fusion-optimists or transhumanists have failed to support their case for enhancement. Indeed, *The Transhumanist Frequently Asked Questions* notes that transhumanists are keenly aware that this issue has been neglected:

While the concept of a soul is not used much in a naturalistic philosophy such as transhumanism, many transhumanists do take an interest in the related problems concerning personal identity (Parfit 1984) and consciousness (Churchland 1988). These problems are being intensely studied by contemporary analytic philosophers, and although some progress has been made, e.g. in Derek Parfit’s work on personal identity, they have still not been resolved to general satisfaction. (Bostrom 2003: section 5.4)

Our discussion also raises general lessons for all parties involved in the enhancement debate, even where purely biological enhancements are concerned. For when one considers the enhancement debate through the lens of the metaphysics of personhood, new dimensions of the debate are appreciated. The literature on the nature of persons is extraordinarily rich, and when one defends or rejects a given enhancement, it is important to determine whether one’s stance on the enhancement in question is truly supported by, or even compatible with leading views on the nature of persons.

Perhaps, alternately, you grow weary of all this metaphysics. You may suspect that social conventions concerning what we commonly consider to be persons are all we have because metaphysical theorizing will never conclusively resolve what persons are. However, as unwieldy as metaphysical issues are, not all conventions are worthy of acceptance, so one needs a manner of determining which conventions should play an important role in the enhancement debate and which ones should not. And it is hard to accomplish this without getting clear on one’s conception of persons. Further, it is difficult to avoid at least implicitly relying on a conception of persons when reflecting on the case for and against enhancement. For what is it that ultimately grounds your decision to enhance or not to enhance, if not that it will somehow improve you? Are you perhaps merely planning for the well‐being of your successor?

**We will return to personal identity in Chapter Eight. There, we will consider a related position on the fundamental nature of mind that says that the mind is software. But let’s pause for a moment. I’d like to raise the ante a bit.** We’ve seen that e**ach of us alive today may be one of the last rungs on the evolutionary ladder that leads from the first living cell to synthetic intelligence. On Earth,** *Homo sapiens* may not be the most intelligent species for that much longer. **In the next chapter, I’d like to explore the evolution of mind in a cosmic context.** The minds on Earth -- past, present and future -- may be but a small speck in the larger space of minds, a space that spans all of spacetime. As I write this, civilizations elsewhere in universe may be having their own singularities.

**Chapter Seven: A Universe of Singularities**

**In your mind’s eye, zoom away from Earth. Envision Earth becoming but a “pale blue dot” in outer space, to use an expression of Carl Sagan’s. Now zoom out of the Milky Way Galaxy. The scale of the universe is truly staggering. We are but one planet in an immense, expanding universe. Astronomers have already discovered thousands of exoplanets, Earthlike planets that seem to have the sort of conditions that led to the development of life on Earth. As we gaze up into the night sky, life could be all around us.**

**This chapter shall illustrate that the technological developments we are witnessing today on Earth may have all happened before, elsewhere in the universe. That is, the universe’s greatest intelligences may be synthetic, having grown out of civilizations that were once biological.**[[48]](#endnote-48) **The transition from biological intelligence to synthetic intelligence may be a general pattern, instantiated over and over, throughout the cosmos. As you read these words, there may be thousands, or even millions, of other worlds that have developed AI technology.**

**If a civilization develops the requisite AI technology, and cultural conditions are favorable, the transition from biological to postbiological may take only a few hundred years. A note on the expression “postbiological.” Consider a biological mind that achieves superintelligence through purely biological enhancements, such as nanotechnologically enhanced neural minicolumns. This creature would be postbiological, although many wouldn’t call it an “AI.” Or consider a computronium that is built out of purely biological materials, like the Cylon Raider in the reimagined *Battlestar Galactica* TV series. Such is artificial, and postbiological.**

**The key point is that there is no reason to expect humans to be the highest form of intelligence out there.** It is humbling to conceive of this, but we may be intellectual lightweights, when viewed on a galactic scale, at least until if, and when, we enhance in radical ways. The gap between us and them could be like that between humans and a a cat, or even a goldfish.

**The Postbiological Cosmos Approach**

Within the field of astrobiology, this position has been called the *postbioligical cosmos* approach. This approach says that the most intelligent alien civilizations will be superintelligent AIs. What is the rationale for this? Three observations, when considered together, motivate this conclusion.

*The short window observation*

Many have urged that once a society creates the technology that could put them in touch with intelligent life on other planets, there is only a short window before they change their own paradigm from biology to AI-- perhaps only a few hundred years.[[49]](#endnote-49) This makes it more likely that the aliens we encounter, if we encounter any, would be postbiological. Indeed, the short-window observation seems to be supported by human cultural evolution, at least thus far. Our first radio signals occurred only about 120 years ago, and space exploration is only about 50 years old, but many Earthlings are already immersed in digital technology, such as cell phones and laptop computers. Currently, billions of dollars are being poured into the development of sophisticated AI, which is now expected to change the face of society within the next several decades.

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

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

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

*The greater age of alien civilizations*

Proponents of SETI (“the Search for Extraterrestrial Intelligence”) have often concluded that alien civilizations would be much older than our own, if they exist. As the former NASA chief historian, Steven Dick, observes: “…all lines of evidence converge on the conclusion that the maximum age of extraterrestrial intelligence would be billions of years, specifically [it] ranges from 1.7 billion to 8 billion years” (Dick, 2013, p. 468). This is not to say that all life evolves into intelligent, technological civilizations. It is just to say that there are much older planets than Earth. Insofar as intelligent, technological life does evolve on even some of them, these alien civilizations are projected to be millions or billions of years older than us, so many could be vastly more intelligent than we are. They would be superintelligent, by our standards. It is humbling to conceive of this, but we may be galactic babies and intellectual lightweights. When viewed on a cosmic scale, Earth is but a playpen for intelligence.

But would the members of these superintelligent civilizations be forms of AI? Even if they were biological, merely having biological brain enhancements, their superintelligence would be reached by artificial means. I suppose some might regard this as “artificial intelligence.” But I suspect something stronger than this, which leads me to my third observation:

*It is likely that these synthetic beings will not be biologically-based.*

We’ve already observed that silicon appears to be a better medium for information processing than the brain itself. In addition, other, superior kinds of microchips are currently under development, such as grapheme and carbon nanotubes. And again, the number of neurons in the human brain is limited by cranial volume and metabolism, but computers can be remotely connected across the globe. AIs can be constructed by reverse engineering the brain, and improving upon its algorithms. And AI is more durable, and can be backed up.

In sum: I’ve 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 superintelligent AIs, because silicon (and likely other materials) are a superior medium for superintelligence. From all this, I conclude that if life is indeed present on many other planets, and if advanced civilizations do tend to develop and then survive, the members of most advanced alien civilizations will likely be superintelligent AIs.

But how does all this mesh with the suggestion of the previous chapter that we humans must bear in mind the difficult, and possibly intractable issues of personal identity when deciding whether to enhance? Intriguingly, if I am right that the greatest intelligences will be postbiological, then the most intelligent civilizations didn’t halt enhancement, if and when they were faced with the kind of concerns that I voiced. Or, perhaps they were unwittingly supplanted by their AI creations, losing control of them.

Perhaps they didn’t halt enhancements because they found some clever solution to these puzzles; or, perhaps their solutions were not so clever, but the solutions led them to enhance in radical ways, nevertheless. Or perhaps, in some world near Alpha Centauri, the aliens didn’t have philosophical discussions. And perhaps on some other distant world they did, but they concluded, based on reflections of alien philosophers who have views akin to the Buddha or Derek Parfit, that there is no real survival. So they opted to upload their minds. As uploads, they were able to quickly develop further enhancements, and in the blink of an eye, they reached superintelligence. Another possibility is that a given alien civilization takes great care in enhancing a given individual during one’s lifetime, so as to not violate certain principles of personal identity, but it uses reproductive technologies to create new members of the species with highly enhanced abilities.

The point here is that the most intelligent civilizations may be the ones that didn’t halt enhancement efforts. This could be due to a variety of reasons. For instance, perhaps they simply haven’t thought about personal identity, or if they have, they do not care about the survival of the self, perhaps because they do not believe in the self at all. They are what the philosopher Pete Mandik called “metaphysically daring,” (Mandik 2015; Schneider and Mandik forthcoming). They are willing to make a bet about whether consciousness or identity of the self can be preserved when one transfers the informational structure of the brain from tissue to silicon chips. Whether these aliens are good philosophers or not, they still reap the benefits. These sorts of beings go ahead and enhance. As Mandik suggests, systems that have high degrees of metaphysical daring could, through making many more digital backups of themselves, be more fit in a Darwinian sense than more cautious beings in other civilizations.(cite Pete Mandick)

In any case, even if I am wrong about all this -- 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 superintelligent AIs. 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 Earthlings first encounter, even if they aren’t the most common. Remember, even the nearest star, Alpha Centauri, is over 4.4 light years away.

The science fiction-like flavor of these issues can encourage misunderstanding, so it is worth stressing that I am not claiming that most life in the universe is non-biological, being AI, contra some news reports of my position. This is absurd. Most life on Earth itself is microbial. Nor am I saying that the universe will be “controlled” or “dominated” by a single superintelligent AI, akin to Skynet, although it is worth reflecting on AI safety in the context of these issues. (Indeed, we shall do so in shortly). I am merely suggesting that the most advanced alien civilizations, if they exist at all, will likely be superintelligent, being vastly older than us, and will likely have become postbiological. Further, I am not saying that these creatures will be made of silicon; for again, candidate alternate substrates to silicon are even under development on Earth, and it is difficult to anticipate what the most efficient substrate is. The point is that they will likely be highly engineered beings: postbiological, enhanced superintelligences.

**Suppose I am right. Suppose that if there is intelligent life out there, the greatest intelligences are AIs. What should we make of this? Here, current debates over AI on Earth are telling. Two important issues—the so-called “control problem” and the nature of mind and consciousness—impact our understanding of what superintelligent alien civilizations may be like. Let’s begin with the control problem.**

**The Control Problem**

We’ve seen that transhumanists as well as advocates of the postbiological cosmos approach in astrobiology suspect that machines will be the next phase in the evolution of intelligence. You and I, how we live and experience life right now, are just an intermediate step, a rung on the evolutionary ladder. These individuals tend to **have an optimistic view of the postbiological phase of evolution. Others, in contrast, are deeply concerned that humans could lose control of superintelligence, however, as it could rewrite its own code and outthink any safeguards we build in. AI could be our final invention. This has been called the “control problem”—the problem of how we Earthlings can control an AI is that is both inscrutable and vastly smarter than us.**

# **We’ve seen that superintelligent AI 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 its designers can no longer predict or understand the technological changes as they unfold. But even if it arises in less dramatic fashion, there may be no way for us to predict or control its goals. Even if we could decide on what moral principles to build into our machines, moral programming is difficult to specify in a foolproof way, and any such programming could be rewritten by a superintelligence in any case. A clever machine could bypass safeguards, such as kill switches, and could potentially pose an existential threat to biological life.**

# The control problem is a serious problem -- perhaps it is even insurmountable. Indeed, upon reading Bostrom’s compelling book on the control problem, *Superintelligence: Paths, Dangers and Strategies*, scientists and business leaders such as Stephen Hawking, Bill Gates, among others, were widely reported by the world media as commenting that superintelligent AI could threaten the human race, having goals that humans can neither predict nor control. **At this time, millions of dollars are pouring into organizations devoted to AI safety. Some of the finest minds in computer science are working on the problem. We shall return to the control problem shortly; there, I will offer some strategies for understanding superintelligence, based on my NASA funded project on this theme. If we can better understand general properties of superintelligent systems, perhaps this will give us a leg up in containing rogue superintelligence on Earth, should the need arise. But for now, let us consider the implications of the control problem on the Search for Extraterrestrial Intelligence (SETI) project.**

**Active SETI**

It is intriguing to consider the control problem from a cosmic perspective. If one takes the control problem seriously, it would be short-sighted to ignore the potential danger that the discovery of alien superintelligence may present. Advocates of Active SETI hold that that, instead of just listening for signs of extraterrestrial intelligence, we should be using our most powerful radio transmitters, such as the giant dish-telescope at Arecibo, Puerto Rico, (pictured below) to send messages in the direction the nearest stars that are nearest to Earth (Shostak, 2015; Falk, 2015).



(Wiki, image search using telescope and location)

**Active SETI strikes me as reckless when one considers the control problem, however. Although a truly advanced civilization would likely have no interest in us, until we have reached the point at which we can be confident that the superintelligent AI(s) does not pose a threat to us, we should not call attention to ourselves, as even one hostile civilization among millions could be catastrophic. Advocates of Active SETI point out that our radar and radio signals are already detectable, but these signals are fairly weak and quickly blend with natural Galactic noise. We would be playing with fire if we transmitted stronger signals that were intended to be heard.**

# **The safest mindset is intellectual humility.** Indeed, barring blaringly obvious scenarios in which alien ships hover over Earth, as in films like *Arrival* and *Independence Day,* I wonder if we could even recognize **the technological markers of a truly advanced superintelligence. Some scientists project that superintelligent AIs could be found near back holes, feeding off their** energy. [[50]](#endnote-50)Alternately, perhaps superintelligences would create Dyson Spheres, megastructures that harnesses the energy of an entire star, such as that pictured below.



# <https://www.sciencealert.com/researchers-just-found-a-second-dyson-sphere-star>

But these are just speculations from the vantage point of our current technology; it’s simply the height of hubris to claim that we can foresee the computational abilities and energy needs of a civilization millions or even billions of years ahead of our own.

Some of the first superintelligent AIs could have cognitive systems that are modeled after biological brains—the way, for instance, that deep learning systems are roughly modeled on the brain’s neural networks. Their computational structure might be comprehensible to us, at least in rough outlines. They may even retain goals that biological beings have, such as reproduction and survival.[[51]](#endnote-51)

But superintelligent AIs, being self-improving, could quickly transition to an unrecognizable form. Perhaps some superintelligences will opt to retain cognitive features that are similar to those of the species they were originally modeled after, placing a design ceiling on their own cognitive architecture. Who knows? But without a ceiling, an alien superintelligence could quickly outpace our ability to make sense of its actions, or even look for it.

An advocate of Active SETI will point out that this is precisely why we should send signals into space—let the superintelligent civilizations locate us, and let them design means of contact they judge to be tangible to an intellectually inferior species like us. While I agree this is reason to consider Active SETI, sadly, the possibility of encountering a dangerous superintelligence outweighs it. For all we know, malicious superintelligences could infect planetary AI systems with viruses, and wise civilizations build cloaking devices. We humans may need to reach our own singularity before embarking upon Active SETI. Our own superintelligent AIs will be able to inform us of the prospects for galactic AI safety and how we would go about recognizing signs of superintelligence elsewhere in the universe. It takes one to know one.

It is natural to wonder whether all this means that humans should avoid developing sophisticated AIs for space exploration; after all, recall the iconic HAL in *2001: A Space Odyssey.* Considering a future ban on AI in space would be premature, I believe. By the time humanity is able to investigate the universe with its own AIs, we humans will likely have reached a tipping point. We will have either already lost control of AI—in which case space projects initiated by humans will not even happen—or achieved a firmer grip on AI safety. Time will tell.

**Superintelligent Minds**

[Note: These two sections are still a bit dense and I will make it more reader friendly while I wait for reviewer comments.]

Raw intelligence is not the only issue to consider. Notice that the postbiological cosmos approach involves a radical shift in our usual perspective about intelligent life in the universe. Normally, we expect that if we encountered advanced alien intelligence we would encounter creatures with very different *biological* features than us. But the postbiological cosmos approach suggests otherwise. To better understand the mind of an advanced intelligence, we must move our focus away from the purely biological case and consider the computational abilities of postbiological beings.

In reflecting on postbiological intelligence, we are not just considering the possibility of alien intelligence -- we may be reflecting upon the nature of ourselves or our descendants as well, for we’ve seen that human intelligence may itself become postbiological. So, in essence, the line between “us” and “them” blurs, as our focus moves away from biology to the difficult task of understanding the computations and behaviors of superintelligence.

This being said, what would the impact on society be, in the event that we learned that vastly more intelligent beings existed elsewhere in the universe, and that they were not even biological? The impact would of course depend on various contextual details that are hard to predict, but it seems fair to say that finding out that a superior intelligence had evolved beyond biological life and become synthetic could be sobering. For we Earthlings would likely ask: is this a general pattern, throughout the cosmos? And is this our future? Would it feel a certain way to be them, from the inside?

The standard view on this latter question is that if we ever encountered advanced alien intelligence, we would encounter biologically distinct creatures, but they would still have minds like ours in an important sense—there would be something it is like, from the inside, to be them. We’ve seen that throughout your daily life, and even when you dream, it feels like something to be you. Likewise, there is also something that it is like to be a biological alien, if such exist—or so we tend to assume. But would a superintelligent AI even have conscious experience and, if it did, could we tell? And how would its inner life, or lack thereof, impact its capacity for empathy and the kind of goals it has?

We considered these issues in detail in earlier chapters, and we can now appreciate their cosmic import. I’ve noted that the question of whether an AI could have an inner life should be key to how we value its existence, for consciousness is central to our judgment of whether it is a self or person. An AI could even be superintelligent, outperforming humans in every cognitive and perceptual domain, but if it doesn’t feel like anything to be the AI, it difficult to view these beings as having the same value as conscious beings, being persons or selves. And conversely, I’ve observed that whether AI is conscious may also be key to how it values *us*. For a conscious AI could recognize in us the capacity for conscious experience.

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

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

So, the issue of machine consciousness is key to how we would react to the discovery of superintelligent aliens. And while I hesitate to speak for world religions, discussions with my colleagues working in astrobiology at the Center of Theological Inquiry suggest that many would reject the possibility that AIs could have souls, or are somehow made in God’s image, if they are not even conscious beings. Pope Francis has recently commented that he would baptize an extraterrestrial (Consolmagno & Mueller, 2014). But I wonder how Pope Francis would react if asked to baptize an AI, let alone one that is not capable of consciousness.

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

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

**Biologically Inspired Superintelligences**

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, at least 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:

**Speed superintelligence** - even a human emulation could in principle run so fast that it could write a PhD thesis in an hour.

**Collective superintelligence** - the individual units need not be superintelligent, but the collective performance of the individuals outstrips human intelligence.

**Quality superintelligence** - at least as fast as human thought, and vastly smarter than humans in virtually every domain (Bostrom 2014).

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 could share. Bostrom’s suggests the following thesis:

The Orthogonality Thesis: “Intelligence and final goals are orthogonal -- more or less any level of intelligence could in principle be combined with more or less any final goal.” (Bostrom 2014, p. 107)

Bostrom is careful to underscore that a great many unthinkable kinds of superintelligences could be developed. At one point, he raises a sobering example of a superintelligence that runs a paper clip factory. Its final goal is the banal task 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’s sobering point is that superintelligence could utilize every form of matter on Earth in support of this goal, wiping out biological life in the process.

The paper clip example illustrates that superintelligence could be of an unpredictable nature, having thinking that is “extremely alien” to us (p. 29). He lays out several ways that a superintelligence could originate. For instance, clever programmers could construct it in unexpected ways, and its architecture may not be derived from the human brain whatsoever. Bostrom 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.

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

The Instrumental Convergence Thesis: “Several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent’s goal being realized for a wide range of final goals and a wide range of situations, implying that these instrumental values are likely to be pursued by a broad spectrum of situated intelligent agents.”(Bostrom 2015, 109)

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

Bostrom does not speculate about superintelligent alien minds in his book, but his discussion is suggestive. Let us call an alien superintelligence that is based on reverse engineering an alien brain, including uploading it, a biologically-inspired superintelligent alien (or “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 superintelligent AIs is highly heterogeneous, with members generally bearing little resemblance to each other. It may turn out that of all superintelligent AIs, 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 its own survival and reproduction, or at least the survival and reproduction of the members of its 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 “mindchildren” 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 wish to change its architecture fundamentally. It may opt for smaller, more gradual improvements, improvements which nevertheless gradually lead the individual toward superintelligence. Perhaps, after reflecting on the personal identity debate, BISAs tend to appreciate the vexing nature of the issues, and they think: perhaps, when I fundamentally alter my architecture, I will no longer be me. Even a being that is an upload, and which believes that it is not identical to the creature that uploaded may nevertheless wish not alter the traits that were most important to their biological counterparts during their biological existence. For remember, uploads are isomorphs, at least at the time they are uploaded, so these are traits that they identify with, at least initially.

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 through connectomics. The field of “connectomics” aims to provide a connectivity map or wiring diagram of the brain, called the “connectome” (Seung 2012).

[insert an image of a connectome and explain it a bit]

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 that are viewpoint invariant. It seems difficult for biologically-based intelligence to evolve without such representations, as they enable categorization and prediction (Hawkins, 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 2011).

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. This has important implications should a SETI program discover a communicating BISA.

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. Clark suggested (1962). I obviously speak to the scenario in which the superintelligent AI’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.

In this chapter, we’ve zoomed away from Earth, situating these issues in a cosmic context. I’ve illustrated that the issues we Earthlings are facing today may not be unique to Earth. In fact, discussions of superintelligence on Earth, together with research in cognitive science, helped inform speculations about what superintelligent alien minds might be like. We’ve also seen that our earlier discussion of synthetic consciousness is relevant as well.

It is also worth noting that as members of these civilizations develop the technology to enhance their own minds, these cultures may confront the same perplexing issues of personal identity that we discussed earlier. Perhaps the most technologically advanced civilizations are the most metaphysically daring ones, as Mandik had suggested. These are the superintelligences that didn’t stall their own enhancements based on concerns about survival. Or perhaps they were concerned about personal identity, and found a clever, or not so clever, way around it.

In what follows, we will descend back to Earth, delving into issues that relate to patternism. It is now time to explore a dominant view of the mind that underlies transhumanism and fusion-optimism. Many transhumanists, philosophers of mind and cognitive scientists have appealed to a conception of the mind in which the mind is software. This is often expressed by the slogan: “the mind is the software the brain runs.” Is this picture of nature of the mind well-founded? If there truly are postbiological intelligences throughout the universe, it is all the more important to consider whether the mind is software.

**Chapter Eight**

**Is the Mind a Software Program?**

[Note: This may not be as refined as the earlier chapters as it is newer.]

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.

-Stephen Hawking[[52]](#endnote-52)

One morning I awoke to a call from a *New York Times* reporter. She wanted to talk about Kim Souzzi, a twenty three year old who had died of brain cancer. A cognitive science major, Kim was planning to become a neuroscientist. But the day she learned she had an exciting new internship, she also learned she had a brain tumor. She posted on Facebook: “Good news: got into The Center for Behavioral Neurosciences’ BRAIN summer program…Bad news: a tumor got into my BRAIN.”[[53]](#endnote-53)

In college, Kim and her boyfriend Josh had shared a common passion for transhumanism. When conventional treatments failed, they turned to cryonics, a medical technique that uses ultra-cold temperatures to preserve the brain upon death. Kim and Josh hoped that she could be revived at some point in the future, when there was a means to do so, as well as a cure for her cancer.

So Kim contacted Alcor, a nonprofit cryopreservation center in Scottsdale, Arizona. She launched a successful online campaign to get the eighty thousand dollars needed for the cryopreservation of her head. To facilitate the best possible cryopreservation, she was advised to spend the last weeks of her life near Alcor.[[54]](#endnote-54) So Kim and Josh moved to a hospice in Scottsdale. In her last weeks, she denied herself food and water to hasten her death, so the tumor would not further ravage her brain.



(Image from https://www.youtube.com/watch?v=wdTdWs4NleA)

(Kim Souzzi is pictured here during an interview at Alcor, next to the containers where she and others are now frozen.)

Cryonics is controversial. Cryopreservation is employed in medicine to maintain human embryos and animal cells for as long as three decades.[[55]](#endnote-55) But when it comes to the brain, cryopreservation is still in its infancy, and it is unknown whether someone cryopreserved using today’s incipient technology could ever be revived. But Kim had weighed the pros and cons carefully.

Sadly, although Kim would never know this, her cryopreservation did not go smoothly.[[56]](#endnote-56) When the medical scans of her brain arrived, they revealed that the cryoprotectant only reached the outer portion of her brain, possibly due to vascular impairment from ischemia, leaving the remainder vulnerable to ice damage.[[57]](#endnote-57) Given the damage, the author of the *New York Times* article, Amy Harmon, considered the suggestion that once uploading technology becomes available, Kim’s brain be uploaded into a computer program. As she noted, certain cryopreservation efforts are turning to uploading as a means of digitally preserving the brain’s neural circuits.

Harmon’s point was that uploading technology might Kim, and cases like hers in which cryopreservation and illness may have damaged too much of the brain for a biological revival. The idea was that, in Kim’s case, the damaged parts of the biological brain could be repaired digitally. That is, the program that her brain was uploaded to could include algorithms carrying out computations that the missing parts were supposed to achieve. And this computer program – this was supposed to be Kim.[[58]](#endnote-58)

Oh boy, I thought. As a mother of a daughter only a few years younger than Kim, that night, I had trouble sleeping. I kept dreaming of Kim. It was bad enough that cancer stole her life. It is one thing to cryopreserve and revive someone. There are scientific obstacles here, and Kim knew the risks. But uploading is another issue entirely. Why see uploading as a means of “revival”?

Kim’s case makes all our abstract talk of radical brain enhancements so much more real. Transhumanism, fusion-optimism, artificial consciousness -- it all sounds so science fiction-like. But in the last chapter, we saw that civilizations throughout the universe may be postbiological. And Kim’s example illustrates that even here on Earth, these ideas are altering lives. Hawking’s remarks voice an understanding of the mind that is in the air nowadays: the view that mind is a program. Philosophers and cognitive scientists express the view with the well-known slogan: “the mind is the software of the brain.” The *New York Times* piece reported that Kim herself had this view of the mind, in fact.[[59]](#endnote-59) If Souzzi, Hawking, and others are right, maybe she can be uploaded after all. So perhaps I should put my earlier doubts about uploading aside.

Chapters Five and Six had urged that uploading is far-fetched, however. It seemed to lack clear support from theories of personal identity. Even modified patternism didn’t support uploading, as modified patternism added the spatiotemporal continuity requirement. To survive uploading, your mind would have to transfer to a new location, outside your brain, through an unusual process in which information about every molecule in your brain is sent to a computer and converted into a software program. Objects that we ordinarily encounter do not “jump” across spacetime to new locations in this way. No single molecule in your brain moves to the computer, but somehow, as if by magic, your mind is supposed to transfer there.[[60]](#endnote-60)

This is perplexing. For this to happen, the mind must be radically unlike ordinary physical objects. My coffee cup is here, next to my laptop; when it moves, it follows a path through spacetime. It isn’t dismantled, measured, and then, across the globe somewhere, configured with new components that mirror its measurements. And if it was, we wouldn’t think it was the same cup, but a replica.

Further, recall the Reduplication Problem. For instance, suppose you try to upload, and consider a scenario in which your brain and body survive the scan, as may be the case with more sophisticated uploading procedures. Suppose your upload is downloaded into an android body that looks just like you, seeming human. Feeling curious, you decide to meet your upload in a bar. As you sip a glass of wine with your android double, the two of you debate who is truly the original – who is truly *you*. Your upload argues convincingly that it is you, for it has all your memories and even remembers the beginning of the surgical procedure in which you were scanned. Your doppelganger even asserts that it is conscious. This may be true. For we saw that if the upload is extremely precise, it may very well have a conscious mental life. But that doesn’t mean it is you. For you are sitting right across from it in the bar.

In addition, if you really uploaded, you would be in principle downloadable to multiple locations at once. Suppose a hundred copies of you are downloaded. You would be *multiply located,* that is, you would be in multiple places at the same time. This is an unusual view of the self. Physical objects can be located in different places at different times, but not at the *same* time. We seem to be objects, albeit of a special kind: we are living, conscious beings. For us to be an exception to the generality about the behavior of macroscopic objects would be stupendous metaphysical luck.[[61]](#endnote-61)

**The Mind as the Software of the Brain**

Such considerations motivated me to resist the siren song of digital immortality, despite my transhumanist views. But what if Hawking and the others are right? What if we *are* lucky because the mind is a kind of software program?

Suppose that Will Castor the scientist developing uploading in the movie *Transcendence* who tried to upload, was presented with the doubts raised in the last section. We tell him that the copy is not the same as the original. It is unlikely that his brain could be destroyed and that the mere information stream, running on various computers, would truly be him. He might offer the following reply:

*The Software Response. Uploading the mind is like uploading software. Software can be uploaded and downloaded across great distances in seconds, and can even be downloaded to multiple locations at once. We are not like ordinary physical objects at all –our minds are instead programs. So if your brain is scanned under ideal conditions, the scanning process copies your neural configuration (your “program” or “informational pattern”). You can survive uploading insofar as your pattern survives.*

The software response draws from the dominant view of the nature of the mind in cognitive science and philosophy of mind. The view that often goes by the slogan: “The mind is the software of the brain” (Block 1995).[[62]](#endnote-62) Let’s call this position “The Software View.” Many fusion-optimists appeal to the Software View, alongside their patternism. For instance, the computer scientist Keith Wiley writes, in response to my view:

The mind is not a physical object at all and therefor properties of physical objects (continual path through space and time) need not apply. The mind is akin to what mathematicians and computer scientists call ‘information,’ for brevity a nonrandom pattern of data… [[63]](#endnote-63)

If your mind is a program it can be uploaded and then downloaded into a series of different kinds of bodies, and it does not need to follow a continuous trajectory through spacetime. This is colorfully depicted in Rudy Rucker’s dystopian novel, *Software*, where the character runs out of money to pay for decent downloads, and out of desperation, dumps his consciousness into a truck. Indeed, perhaps an upload wouldn’t even need to be downloaded at all. Perhaps it can just reside somewhere in a computer simulation, as in the classic film, *The Matrix*, in which many beings, including the notorious Smith villain, live in the Matrix – a massive computer simulation. Smith is a particularly powerful software program. Not only can he appear anywhere in the Matrix, in pursuit of the good guys, he can be in multiple locations at once. At various points in the movie, Neo even finds himself fighting hundreds of Smiths.

As these science fiction stories illustrate, the Software View is an outgrowth of the age of the Internet. Indeed, elaborations of it even describe the mind with expressions like “downloads” “apps” and “files.” As Steven Mazie at *Big Think* puts it:

Presumably you’d want to Dropbox your brain file (yes, you’ll need to buy more storage) to avoid death by hard-drive crash. But with suitable backups, you, or an electronic version of you, could go on living forever, or at least for a very, very long time, “untethered,” as Dr. Schneider puts it, “from a body that’s inevitably going to die. (Mazie 2014)

Another proponent of patternism is the neuroscientist and head of the *Brain Preservation Foundation*, Ken Hayworth, who is irked by my critique of patternism. For it is apparently really obvious that the mind is a program:

It always boggles my mind that smart people continue to fall into this philosophical trap. If we were discussing copying the software and memory of one robot (say R2D2) and putting it into a new robot body would we be philosophically concerned about whether it was the ‘same’ robot? Of course not, just as we don’t worry about copying our data and programs from an old laptop to a new one. If we have two laptops with the same data and software do we ask if one can ‘magically’ access the other’s RAM? Of course not.[[64]](#endnote-64)

So, is the Software View correct? Is this what the mind is? If so, the software response is onto something. Perhaps you can survive radical brain enhancements, and can even upload. As I’ll explain, the software approach to the mind is deeply mistaken. It is one thing to say that the brain is computational; this is a research paradigm in cognitive science that I am quite fond of, (see, for instance, my earlier book, *The Language of Thought*). Although SAM is often taken as being part and parcel of the computational approach to the brain, there are many metaphysical approaches to the nature of mind that are compatible with a computational approach to the brain (Schneider 2011a). And, as I’ll explain shortly, the view that the mind or self is software is one we should do without.

Before we assess the Software View, let’s me say a bit more about its significance. There are at least two reasons why the issue is important. First, if the Software View is correct, patternism is more plausible than Chapters Five and Six indicated, for my objections involving spatiotemporal discontinuity and reduplication can be dismissed. This remove all objections to patternism, however. Recall that it seemed unclear when an alteration in a pattern is compatible with survival and when it is not. But it encourages one to take patternism seriously, despite its flaws. For if the mind is software, perhaps there is some sort of pattern that is compatible with survival, even if, from an epistemic standpoint, it is tricky to find out what that underlying pattern is. This epistemological issue still makes enhancement decisions difficult, but if the Software View is right, we should view uploading with less skepticism.

Second, if the Software View is correct, it would be an exciting discovery because it would provide an account of the nature of the mind. Theories of the fundamental nature of mind, like theories of personal identity, were an outgrowth of metaphysics. About fifty years ago, the study of the nature of the mind blossomed into a field of its own, called, “philosophy of mind.” Two central problems lie at the heart of philosophy of mind. The first is the aforementioned hard problem of consciousness. Recall that the hard problem of consciousness asks: why does all this sophisticated information processing need to feel like anything, from the inside? Why do we need to have experience? The second is a problem which we will now discuss, called “the mind-body problem.”

**The Mind-Body Problem**

Suppose that you are sitting in a cafe studying right before a big presentation. All in one moment, you taste the espresso you sip, feel a pang of anxiety, consider an idea, and hear the scream of the espresso machine. What is the nature of these thoughts? Are they just a matter of physical states of your brain, or are they something more? Relatedly, what is the nature of your mind? Is your mind just a physical thing, or is it something above and beyond the configuration of particles in your brain?

These questions pose the Mind-Body Problem. The problem aims to situate mentality within the world that science investigates. The Mind-Body Problem is closely related to the Hard Problem. But the focus of the Hard Problem is consciousness, whereas the Mind-Body Problem focuses on mental states more generally, even nonconscious mental states, and instead of asking why these states must exist, it seeks to determine how they relate to what science investigates.

Contemporary debates over the mind-body problem were launched over fifty years ago, but some classic positions began to emerge as early as the pre-Socratic Greeks. The problem is not getting any easier. There are some fascinating solutions, to be sure. But as with the debate over personal identity, there are no uncontroversial ones in sight. So, does the Software View solve this classic philosophical problem? Let’s consider some influential positions on the problem, and see how the Software View compares.

*1. Panpsychism*

Recall that panpsychism holds that even the smallest layers of reality have experience. Fundamental particles have minute levels of consciousness, and in a watered-down sense, they are subjects of experience. When particles are in extremely sophisticated configurations, such as when they are in nervous systems, more sophisticated, recognizable, forms of consciousness arise. Panpsychism may seem outlandish, but the panpsychist would respond that their theory actually meshes with fundamental physics, for experience is the underlying nature of the properties that physics identifies.

*2. Substance Dualism*

According to this view, reality consists in two kinds of substances, physical things (e.g., brains, rocks, bodies) and nonphysical ones (i.e., minds, selves or souls). While you personally may reject the view that there’s an immaterial mind or soul, science alone cannot rule it out. The most influential philosophical substance dualist, Rene Descartes, thought that the workings of the nonphysical mind corresponded with the workings of the brain, at least during one’s lifetime. Contemporary substance dualists offer sophisticated nontheistic positions, as well as intriguing and equally sophisticated theistic ones.

*3. Physicalism (or “materialism”)*

We discussed physicalism briefly in Chapter Five. According to physicalism, the mind, like the rest of reality, is physical. For everything is either made up of something that physics describes, or it is a fundamental property, law or substance figuring in a physical theory. (Here, by “physical theory” physicalists tend to gesture towards the content of the final theory of everything that a completed physics uncovers, whatever that is.) There are no immaterial minds, and all of our thoughts are ultimately just physical phenomena. This position has been called “materialism,” but it is now more commonly called ‘physicalism.’ Because there is no second immaterial realm, like substance dualism claimed, the view is regarded as a form of monism. For it claims that there is one fundamental type of category to reality: physical entities.

*4. Property Dualism*

The point of departure for this position is the hard problem of consciousness. Proponents of property dualism believe that consciousness is a fundamental feature of the brain, and that it is not reducible to mere physical features of the brain. Consciousness is special, and it defies an explanation that just appeals to neuroscience or physics. Like substance dualism, property dualism holds that reality has two distinct realms. But it rejects souls and immaterial minds. Brains, and perhaps other thinking systems, if such exist, are the things that have the nonphysical properties (or features). These nonphysical features are basic building blocks of reality, but unlike panpsychism, they are not microscopic. They are features of complex systems, i.e., brains.

There are many intriguing approaches to the nature of mind, but I’ve focused on the most influential. Should the reader wish to consider solutions to the problem in more detail, there are several excellent introductions available (cite, Heil, Lowe, Kim, etc.). Now that we’ve considered these positions, let us turn back to the Software View.

**Assessing the Software View**

The Software View has two initial flaws, both of which can be remedied, I believe. First, not all programs are the sorts of things that have minds. For instance, I doubt that the Amazon or Facebook app on your smartphone has a mind. If minds are programs, they are programs of a very special sort, having layers of complexity that fields like psychology and neuroscience find challenging to describe. Second, we’ve seen that consciousness is at the heart of our mental lives. If you sympathize with this, then a zombie program – a program incapable of having experience – just isn’t the sort of thing that has a mind.

These points are not decisive objections, for if proponents of the software view agree with one or both of these criticisms, they can restrict their view. For instance, if they agree with both criticisms, they can restrict the Software View in the following way:

Minds are programs of a highly sophisticated sort, which are capable of having conscious experiences.

The problems I will subsequently raise with the Software View will apply to both versions of the view, so we will not need to differentiate versions of the Software View in the rest of this chapter. I’ll simply refer to all versions as “The Software View.”

To determine whether The Software View is plausible, let us ask: what is a program? As the image below indicates, a program is a list of instructions in lines of code. The lines of code are instructions in a programming language that tell the computer what tasks to do. Most computers can execute several programs, and in this way, new capacities can be added or deleted from the computer.



A line of code is like a mathematical equation. It is highly abstract, standing in stark contrast with the concrete physical world around you. You can throw a rock. You can lift a coffee cup. But just try to throw an equation. You can’t do it, of course. This is because equations are abstract entities; they are not situated in space or time.

Now that we appreciate that a program is an abstract entity, we can see a serious flaw in the Software View. If your mind is a program, then it is just a long sequence of instructions in a programming code. The Program View is saying the mind is an abstract entity. But think about what this means. The field of philosophy of mathematics studies the nature of abstract entities -- entities like equations, sets and programs. Abstract entities are said to be *non-concrete: they are non-spatial, nontemporal, non-physical and acausal.* The inscription “5” is here on this page, but the actual number, as opposed to the inscription, isn’t located anywhere. Abstract entities are not located in space or time, they are not physical objects, and they do not cause events to occur in the spatiotemporal manifold.

How can the mind be an abstract entity, like an equation or the number two? It is a category mistake to think minds (or selves and persons) are equations or programs. We are spatial beings. Further, our minds have states that cause us to act in the world. And moments pass for us. We are temporal beings. So, your mind is not an abstract entity like a program. And there is still reason to doubt that uploading the mind is a true means for Kim Souzzi, or others, to survive.

As I’ve been stressing throughout the second part of this book, assuming that there are selves that persist over time, biologically-based enhancements that gradually restore and cautiously enhance the functioning of the biological brain are a safer route to longevity and enhanced mental abilities, even if the technology to upload a complete human brain is developed. Fusion-optimists tend to endorse both rapid alterations in psychological continuity or changes in one’s substrate. Both types of enhancements seem risky, at least if one believes that there is such a thing as a persisting self.

I emphasized this in Chapter Five and Six as well, although there, my rationale did not concern the abstract nature of the Software View. There, my caution stemmed from the controversy in metaphysics over which, if any, competing theories of the nature of the person are correct. This left us adrift concerning whether radical, or even moderate, enhancements are compatible with survival. We can now see that just as patternism about the survival of the person is flawed, so too, the related Software View is problematic.

I’d like to caution against drawing a certain conclusion from my rejection of the Software View, however. As I’ve indicated, the computational approach to the mind in cognitive science is an excellent explanatory framework (see Schneider, 2011b, for a defense). But explanation in cognitive science does not entail the view that the mind is a program. Consider Ned Block’s canonical paper, “The Mind as the Software of the Brain” (1995). Aside from its title, which I obviously disagree with, it astutely details many key facets of the view that the brain is computational: e.g., cognitive capacities such as intelligence and working memory are explainable via the method of functional decomposition, mental states are multiply realizable, the brain is a syntactic engine driving a semantic one. Block is accurately describing an explanatory framework in cognitive science by isolating key features of the computational approach to the mind. None of this entails the metaphysical position that the mind is a program, however.

The Software View wasn’t a viable position then. But you might wonder if the transhumanist or fusion-optimist could provide a more feasible computationalist approach to the nature of the mind. As it happens, I do have a further suggestion. To introduce this, I would like to take a bit of a detour, one that will ultimately lead us to formulate a different, transhumanist-inspired, view of the mind. On this view, minds are program instantiations.

**Could Commander Data be Immortal?**

Consider Commander Data from *Star Trek: the Next Generation.* Suppose he finds himself in an unlucky predicament, on a hostile planet, surrounded by aliens that are about to dismantle him. In a last-ditch act of desperation, he quickly uploads his artificial brain onto a computer on the *Enterprise*. Does he survive? And could he, in principle, do this every time he’s in a jam, so that he’d be immortal?

If I’m correct that the mind is not a software program, this has bearing on the question of whether A.Is, including uploads, could achieve immortality, or, rather, whether they could achieve what we might call “functional immortality.” Notice that I write “functional immortality” because the universe will eventually have a heat death. Nothing in spacetime can be truly immortal, unless, that is, some superintelligent civilization finds a cure for the heat death of the universe! But I’ll ignore this technicality in what follows.

The Software View drew from a commonly held view of machine minds: the idea that a conscious android or other kind of AI could achieve functional immortality by creating backup copies of itself, and thus transfer its consciousness from one computer to the next when an accident happens. But I think this cultural view is mistaken – androids can’t be functionally immortal either. Just as a particular biological mind is a concrete entity, and thus not a program, so too, a particular A.I. mind is not a program; although it may be, in a related vien, the concrete entity running a program. A particular A.I. is vulnerable to destruction by accident or the slow decay of its parts, just as we are, at least if its cognitive and perceptual systems are housed in the same location as its body. And if its artificial brain is housed somewhere else, that spot is vulnerable. (Indeed, if I am right, then perhaps superintelligent AI’s, if they exist, squirrel themselves away in isolated pockets of the universe, guarding their locations carefully, and they are surrounded by cloaking devices!)

Perhaps the confusion about A.I. and uploads being immortal arises from the fact that discussions of tend to be ambiguous as to whether “A.I.” refers to a particular A.I. (i.e., an individual being) or to a *type* of A.I. system (i.e., an abstract entity). It is important to disambiguate claims about survival. Perhaps, if one wants to speak of types of programs as “types of mind,” the types could be said to “survive” uploading, according to two watered-down notions of survival. First, perhaps, at least in principle, a machine that contains a high fidelity copy of an uploaded brain, can, at least in principle, run the same program as a human brain did before it was destroyed by the uploading procedure. Some may draw from this that the type of mind “survives,” although I find this to be a very watered-down notion of survival because, as our discussion of uploading emphasized, no single conscious being persists.

Second, a program, as an abstract entity, is timeless. It does not cease to exist because it is not a temporal being. (This is not “survival” in a serious sense, but it is worth mentioning, as *not ceasing to exist* is a property that is necessary for survival.) This may lead one to believe that the AI survives, especially if one does not distinguish between types and tokens. Particular selves or minds do not survive in either of these two senses.

This is all highly abstract. So let’s go back to the example of Commander Data. Data is a particular A.I., and as such, he is vulnerable to destruction. There may be other androids of this type, (individual A.I.s themselves), but their survival does not ensure the survival of Data, it just ensures the “survival” of Data’s type of mind. (Herein, I write “survival” in scare quotes to indicate that I am referring to the aforementioned watered-down sense of survival.)

So there Data is, on a hostile planet, surrounded by aliens that are about to destroy him. So Data quickly uploads his artificial brain onto a computer on the *Enterprise*. Does he survive or not? On my view, we now have a distinct token of the type of mind, *Data*, being run by that particular computer. We could ask: Can that token survive the destruction of the computer by uploading again, i.e., transferring the mind of that token to a different computer? No. Again, uploading would merely create a different token of the same type. An individual’s survival depends on where things stand at the token level, not at the level of types.

However, it is worth underscoring that a *particular* A.I. could live a very long time, insofar as its parts are extremely durable. Perhaps Data could even achieve functional immortality by avoiding accidents and having his parts replaced as they wear out. My view is compatible with this scenario, for Data’s survival in this case does not happen by transferring his program from one physical object to another. On the assumption that one is willing to grant that humans survive the gradual replacement of their parts over time, why not also grant it in the case of A.I.s? Of course, in Chapter Five, I emphasized that it is controversial whether persons survive the replacement of their parts; perhaps the self is an illusion, as Derek Parfit, Fredrick Nietsche and the Buddha have suggested.

**Is Your Mind a Program Instantiation?**

Now let us turn back to the Software View. Throughout my discussion of Data, notice that I’ve been implicitly discussing a view of the mind in which the mind is a program *instantiation* (a concrete thing), and in which *types* of minds are programs. Although I’ve rejected the view that the mind is software, for the purpose of illustration, in these few paragraphs I have introduced this related view.

What if the proponent of the Software View held that the mind is not the program but the *instantiation* of the program -- the thing that runs the program or has the informational pattern? Unlike equations and programs, program instantiations are concrete spatiotemporal objects, what metaphysicians call “tokens.” Consider some properties of things around you: blueness, flatness, and so on. Property *tokens*, like the blueness of the sky at a particular time, are spatiotemporal entities. (In contrast, property types themselves -- the general types of properties that many things can have -- are not tokens, and they don’t exist in spacetime. In this respect, property types are akin to an equation or a program.)

In a similar vein, something that instantiates a program exists in space and time. It is a concrete object. Roughly, something *instantiates* a program when it has a certain pattern of matter and energy – a whir of activity – that is aptly described as running a certain program. The pattern of matter and energy has a structure, and the elements of the structure correspond, in nontrivial ways, to elements of the program (e.g., variables, constants), and further, these elements seem to behave in accordance with the program operations.[[65]](#endnote-65)

Let us call this position *The Software Instantiation* *View*. Importantly, on this view, the mind is not the program or software of the brain – instead, it is the *thing* running the program, whatever that is. So this is *not* a view that is accurately expressed by the slogan: “the mind is the software of the brain.” Remember, we’ve rejected this view. Instead, according to the Software Instantiation View, the mind is a concrete entity, not an abstract entity like a program. From a metaphysical standpoint, this is a major distinction. To summarize this view, it says:

**The Software Instantiation View of the Mind (“SIM” for short)**

(SIM) The mind is the entity running the program, that is, the algorithm that the brain implements, something in principle discoverable by cognitive science.

To see how different SIM is from the Software View, notice that SIM doesn’t suggest you can survive uploading. Again, objects running programs are concrete entities, being part of the spatiotemporal world. If minds are instantiations, my earlier concerns involving spatiotemporal discontinuities still apply to uploading. As with Modified Patternism, each upload or download is not the same person as the original, although it has the same program as the original at the time of its creation.

Of course, it is still important to ask whether, SIM is an insightful approach to the nature of the mind, especially given the popularity of computational approaches to the mind in cognitive science. So how would SIM fare as a response to the mind-body problem? There is an initial problem with SIM, but I believe we can remedy it. As it stands, most advocates of a software instantiation view would reject SIM, as they tend to hold that thinking is *multiply realizable*, being realizable by different substrates, such as silicon-based computers and carbon-based brains.[[66]](#endnote-66)

Technically, SIM says the program is implemented by the *brain*, but both transhumanists and traditional proponents of computationalism generally do not want to limit instantiations to brains, for they hold that other creatures (aliens, A.I.s) could in principle have minds and mental states. Perhaps, there are organic intelligences elsewhere in the universe, or perhaps, as the previous chapter urged, the most intelligent creatures in the cosmos are forms of superintelligent AI.

To fix the problem, let’s modify SIM in the following way:

**The Software Instantiation Approach to the Mind (SIM\*)**

The mind is the entity running the program, that is, the algorithm that the brain or other cognitive system implements, something in principle discoverable by cognitive science.

SIM\*, unlike the original Software View, avoids the category mistake of viewing the mind as abstract. But like the original view and the related patternist position, it draws from computational approaches to the brain in cognitive science.

**SIM\* as an Account of the Mind’s Nature**

So, how does SIM\* fare as an approach to the fundamental nature of mind? Unfortunately, SIM\* is threadbare, telling us little about the nature of the mind. For consider the different approaches to the nature of mind that we discussed earlier. Does SIM\* rule any of them out? If it doesn’t, it isn’t much of a theory of the nature of mind. For instance, consider *panpsychism*. Is that which instantiates the program a system that has fundamental elements that are experiential? SIM\* doesn’t say.

Now consider substance dualism. Consider, for instance, the influential substance dualism developed by Rene Descartes, called *Cartesian dualism*. According to this view, reality consists in two kinds of substances: physical things (e.g., brains, rocks, bodies) and nonphysical ones (i.e., minds, or souls). Why can’t Cartesian minds run programs? After all, is the fact that cognitive science suggests that the brain is computational logically inconsistent with there being a soul? Descartes himself thought that the workings of the nonphysical mind corresponded with the workings of the brain, at least during one’s lifetime. While contemporary philosophers are skeptical of Cartesian dualism, holding that he had an implausible view of the relationship between the mind and brain, my point is simply that SIM\* doesn’t tell us whether Cartesian dualism is true.

Similarly, SIM\* is compatible with physicalism, the view that everything is either made of something that physics describes, or it is a fundamental property, law or substance figuring in a physical theory.[[67]](#endnote-67) And it is compatible with property dualism, for a program can be run by physical systems that have nonphysical properties.

In essence, although (SIM\*) does not venture the implausible claim that the mind is abstract, like the Software View did, it is metaphysically threadbare. For it doesn’t preclude a number of very distinct positions about the nature of mind. It tells us little about the nature of mind, except that it is something that runs a program. And anything could do that, in principle—Cartesian minds, systems made of fundamental experiential properties, and so on.

**Returning to Alcor**

Three years after Kim’s death, Josh gathered her special belongings and returned to Alcor (cite Harmon, NYT). He was making good on a promise to her, delivering her things where she can find them if her brain is revived. Frankly, I wish the conclusion of this chapter had been different. If the Software View was correct, then, at least in principle, minds could be the sort of thing that can be uploaded, downloaded, and rebooted. This would an afterlife for the brain, if you will – a way in which a mind, like Kim’s, could survive the death of the brain. Yet our reflections revealed that The Software View turned minds into abstract objects. So we turned to a related view, in which the mind is a program instantiation. The Program Instantiation View does not support uploading either, and while it is an interesting approach, it is too threadbare, from a metaphysical standpoint, to be much of approach to the nature of mind.

Although I do not have access to the medical details of Kim’s cryopreservation, I do find the *New York Times* report that there was imaging evidence that the outer layers of Kim’s brain was successfully cryopreserved to be a bit hopeful, for as the author noted, the brain’s neocortex seems central to who we are, being key to memory and language (cite). So perhaps a biologically-based reconstruction of the damaged parts would be compatible with the survival of the person. For instance, I’ve observed that even today, there is active work on hippocampal prosthetics; perhaps parts of the brain like the hippocampus are rather generic, and replacing them with biological or even AI-based prosthetics doesn’t change who one is. While one can hold out hope, I’ve emphasized that there is tremendous uncertainty here, due to the perplexing and controversial nature of the personal identity debate. But individuals who were cryopreserved in these early attempts are not akin to the hypothetical patients in the Mind Design Center thought experiment. They were not ordinary people who are out on a shopping trip, and who, being well, have the luxury of deciding against certain enhancements, in order to take a metaphysically humble approach. They were terminally ill patients, who, in need of a cure, agreed to a moonshot treatment.

Does all this mean that uploading projects should be scrapped? Perhaps uploading technology can nevertheless benefit our species. For instance, global catastrophe may make Earth inhospitable to biological life forms, and uploading may be a way to preserve the human way of life and thinking, if not the actual humans themselves. And if these uploads are indeed conscious, this could be something that members of our species come to value, when confronted with their own extinction.

Further, even if uploads aren’t conscious, the use of simulated human minds for space travel could be a safer, more efficient way of sending intelligent beings into space than sending a biological human. The public tends to find manned missions to space exciting, even when robotic missions seem more efficient. Perhaps the use of uploaded minds would excite the public. Perhaps these uploads could even run terraforming operations on inhospitable worlds, readying the terrain for biological humans. You never know.

Further, brain uploading could facilitate the development of brain therapies and enhancements that could benefit humans or nonhuman animals, for uploading part or all of a brain could help generate a working emulation of a biological brain that we could learn from. AGI developers may find it a useful means of AI development. Who knows, perhaps A.I. that is descended from us will have greater chance of being benevolent toward us.

Finally, some humans will understandably want digital backups of themselves. For what if you found out that you were going to die soon. If you felt you couldn’t truly survive uploading, would you wish to make an android double nevertheless? Perhaps you may wish to leave a copy of yourself to communicate with your children or complete projects that you care about.

Indeed, the personal assistants – the Siris and Alexas the future -- might be uploaded copies of deceased humans we have loved deeply. Perhaps people will wish that their descents would be able to converse with something much like them. Perhaps even, our friends will be copies of ourselves, tweaked in ways we find insightful.

**The Afterlife of the Brain**

At the heart of this book was a dialogue between philosophy and science. We’ve seen that science of emerging technologies can challenge and expand our philosophical understanding of the mind, self and person. Conversely, philosophy sharpens our sense of what these emerging technologies can achieve: whether there can be conscious robots, whether you could replace much of your brain with microchips and be confident that it would still be you, and so on.

This book has attempted, very provisionally, to explore mind design space. Although we do not know whether there will be places like *Immortex* or the *Center for Mind Design,* I wouldn’t be surprised. For the events of today speak for themselves: this is a time when AI is projected to replace most blue and white collar jobs over the next several decades, and in which influential figures are suggesting that humans should merge with machines. But I’ve stressed that the future of minds, whether they be human minds or AI minds, is a matter that requires public dialogue and philosophical contemplation. All stakeholders must be involved.

I’ve suggested that it is not a forgone conclusion that sophisticated AIs will be conscious. Instead, I’ve argued for a middle of the road approach, one that rejects the Chinese Room, yet which doesn’t assume, based on the computational nature of the brain or the conceptual possibility of isomorphs, that sophisticated AI will be conscious. Conscious AI may not, in practice, be built, or it may not be compatible with the laws of physics to create consciousness in a different, non-biological, substrate. But by carefully gauging whether the AIs we develop are conscious, we can approach the issue with care. And by publicly debating all this in a way that moves beyond both technophobia and or a tendency to view the AIs that look superficially humanlike as conscious, we will be better able to judge whether and how conscious AI should be built. These can be choices societies make carefully.

I’ve further emphasized that from an ethical standpoint it is best to assume that a sophisticated AI may be conscious, at least until we develop tests for consciousness that we can be confident in. For any mistake could wrongly influence the debate over whether AI might be worthy of special ethical consideration as sentient beings. Not only could a mistake cause needless pain and suffering to sentient beings, but, as films like *Ex Machina* and *I,Robot* illustrate, any failure to be charitable to AI may come back to haunt us, as they may treat us as we treated them.

Some of the younger readers may one day be faced with an opportunity to make mind design decisions. If you are one of these readers, my message to you is this: before you enhance, you must reflect on who you are. If you are uncertain as to the ultimate nature of the person, as I am, take a safe, cautious route: as much as possible, stick to gradual, biologically-based therapies and enhancements, ones that mirror the sort of changes that normal brains undergo in the process of learning and maturation. Bearing in mind all the thought experiments that we discussed that questioned more radical approaches, and the general lack of agreement in the personal identity debate, this cautious approach is most prudent. It is best to avoid radical, quick changes, even one’s that do not alter the type of substrate on has (e.g., carbon versus silicon). For recalling the psychological continuity view, remember that dramatic enhancements might be too much of a shift in one’s mental life. Further, we saw that it is also prudent to avoid attempts to “migrate” the mind to another substrate.

Further, until we know more about synthetic consciousness, we cannot be confident that transferring key mental functions to AI components is safe in parts of the brain underlying consciousness. Of course, we have yet to determine whether AI is conscious, so we do not know whether you, or perhaps more precisely, the AI copy of you, would even be a conscious being, if you tried to merge with AI. If it wouldn’t be conscious, I doubt this creature would really be you.

By now, you can see that any trip to the Center for Mind Design could be vexing, even perilous. I wish I could provide you with a clear, uncontroversial path to guide you through the mind design decisions. Instead, my message has been: as you consider such scenarios, remember to approach the issues with metaphysical humility.

**Appendix One: Transhumanism**

Transhumanism is not a monolithic ideology, but it does have an official declaration and an organization. The World Transhumanist Association is an international nonprofit organization founded in 1998 by philosophers David Pearce and Nick Bostrom. The main tenets of transhumanism were stated in the Transhumanist Declaration (World Transhumanist Association, 1998), which is reprinted below:

*The Transhumanist Declaration*

1. Humanity will be radically changed by technology in the future. We foresee the feasibility of redesigning the human condition, including such parameters as the inevitability of aging, limitations on human and artificial intellects, unchosen psychology, suffering, and our confinement to the planet earth.

2. Systematic research should be put into understanding these coming developments and their long-term consequences.

3. Transhumanists think that by being generally open and embracing of new technology we have a better chance of turning it to our advantage than if we try to ban or prohibit it.

4. Transhumanists advocate the moral right for those who so wish to use technology to extend their mental and physical (including reproductive) capacities and to improve their control over their own lives. We seek personal growth beyond our current biological limitations.

5. In planning for the future, it is mandatory to take into account the prospect of dramatic progress in technological capabilities. It would be tragic if the potential benefits failed to materialize because of technophobia and unnecessary prohibitions. On the other hand, it would also be tragic if intelligent life went extinct because of some disaster or war involving advanced technologies.

6. We need to create forums where people can rationally debate what needs to be done, and a social order where responsible decisions can be implemented.

7. Transhumanism advocates the well-being of all sentience (whether in artificial intellects, humans, posthumans, or non-human animals) and encompasses many principles of modern humanism. Transhumanism does not support any particular party, politician or political platform.

This document was followed by the much longer and extremely informative Transhumanist Frequently Asked Questions, which is widely available online (Cite).

**Bibliography**

Hughes, J. (2013), in The Transhumanist Reader, More, M. and More, N., Boston: Wiley.

Bryson, J., “A Role for Consciousness in Action Selection,” *International Journal of Machine Consciousness* Vol. 4, No. 2 (2012) 471-482

Baars, B. 2008. “The Global Workspace Theory of Consciousness” in Velmans and Schneider, *The Blackwell Companion to Consciousness*. Boston, Mass: Wiley-Blackwell.

Block, N. 1995. “The mind as the software of the brain.” In D. Osherson, L. Gleitman, S. Kosslyn, E. Smith, & S. Sternberg (Eds.), *An Invitation to Cognitive Science* (pp. 377– 421. New York: MIT Press.

Bostrom, N. 2003. “The Transhumanist Frequently Asked Questions”: v 2.1. World Transhumanist Association. Retrieved from http://transhumanism.org/index.php/WTA/faq/.

---- 2005. “History of Transhumanist Thought.” *Journal of Evolution and Technology*, 14(1).

---- 2008. “Dignity and enhancement”. in *The Presidents Council on Bioethics, Human Dignity and Bioethics: Essays Commissioned by the President's Council on Bioethics*, Washington, DC: U.S. Government Printing Office.

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*, vol. 64, pp. 156- 165.

Chalmers, D. 2008. “The Hard Problem of Consciousness” in Velmans and Schneider, *The Blackwell Companion to Consciousness*. Boston, Mass: Wiley-Blackwell.

--- Chalmers, D. (2010). "[The Singularity: A Philosophical Analysis](http://commonsenseatheism.com/wp-content/uploads/2011/01/Chalmers-The-Singularity-a-philosophical-analysis.pdf)." Journal of Consciousness Studies 17:7-65.

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.* NY: Harper and Row.

Cole, David. 2014. "The Chinese Room Argument", *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2014/entries/chinese-room/>.

Corabi, J. and Schneider, S. “The Metaphysics of Uploading,” *Journal of Consciousness Studies* 19 (7):26 (2012)

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

Dick, S. 2013. “Bringing Culture to Cosmos: the Postbiological Universe”, in *Cosmos and Culture: Cultural Evolution in a Cosmic Context*, Dick, S. and Lupisella, M. eds. 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: Doubleday.

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

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: Times Books.

Harmon, A., “A Dying Young Woman’s Hope in Cryonics and a Future”, *New York Times* September 12, 2015

Kurzweil, R. 2005. *The singularity is near: When humans transcend biology*. New York: Viking.

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

Pigliucci, M. 2014. TITLE? in *Intelligence Unbound*, Blackford, R., Broderik, D. (eds.) Boston: Wiley-Blackwell.

Sandberg, A., Boström, N. 2008. “Whole Brain Emulation: A Roadmap.” Technical Report #2008**-**3. Future of Humanity Institute, Oxford University.

Schwitzgebel, E. (2007). *Describing Inner Experience? Proponent Meets Skeptic*.

Schneider, Susan, ed. 2009. *Science Fiction and Philosophy*. Chichester, UK: Wiley- Blackwell.

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

----- 2011b. *The Language of Thought: a New Philosophical Direction*. Boston: MIT Press.

----2014. “The Philosophy of ‘Her’”, *The New York Times*, 2 March. 15

“Superintelligent AI and the Postbiological Cosmos Approach” in Lorsch, A. What is Life? Cambridge: Cambridge University Press, 2017

“Alien Minds,” Discovery (an astrophysics trade anthology, based on a NASA/Library of Congress Symposium), Steven Dick, Cambridge University Press, 2015.

“Testing for Consciousness in AI and Biological Systems,” Liao, M. and Chalmers, D., AI, Oxford: Oxford University Press, forthcoming.

“Artificial Intelligence, Consciousness, and Moral Status,” The Routledge Handbook of Neuroethics, Syd Johnson and Karen Rommelfanger, eds., 2017.

“What Breathes Fire into the Equations? Response to Critics,” Journal of Consciousness Studies, forthcoming in 2017.

“Non-reductive Physicalism Cannot Appeal to Token Identity,” Philosophy and Phenomenological Research, 85 (3):719-728, 2013.

“The Metaphysics of Uploading”, (with Joseph Corabi), Symposium Contribution on David Chalmers’, “The Singularity”, with Chalmers’ response, Journal of Consciousness Studies, 2012.

--Reprinted in Intelligent Machines, Uploaded Minds, Russell Blackford (ed.), Wiley-Blackwell, 2014. (Includes new response to David Chalmers’ reply to the paper, simplified for multidisciplinary audience.)

----“How Philosophy of Mind can Shape the Future,” in Philosophy of Mind in the 20th and 21th Century, Amy Kind (ed.), forthcoming with Routledge, 2016. (The final chapter of a four volume set on the history of philosophy of mind, with Pete Mandik.)

----“The Language of Thought.” The Routledge Companion to Philosophy of Psychology, Paco Calvo and John Symons, eds. NY: Routledge, 2009, pp. 280-295.

----“Thought Experiments: Science Fiction as a Window into Philosophical Puzzles,” in Science Fiction and Philosophy, Susan Schneider, editor. Oxford: Blackwell Publishing, 2009, pp. 1-14.

----“Mindscan: Transcending and Enhancing the Human Brain,” Science Fiction and Philosophy, Susan Schneider, editor. Oxford: Blackwell Publishing, 2009, pp. 241-255.

----“Cognitive Enhancement and the Nature of Persons.” The University of Pennsylvania Bioethics Reader, Art Caplan and Vardit Radvisky, eds., Springer, 2009, pp. 844-856.

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

Searle, J. 2008. “Biological Naturalism”, in *The Blackwell Companion to Consciousness*. Max Velmans and Susan Schneider, eds., Mass: Wiley-Blackwell.

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

Shostak, S. 2009. *Confessions of an Alien Hunter*, New York: National Geographic. Tonini, G. 2008. “The Information Integration Theory of Consciousness” in Velmans and

Schneider, *The Blackwell Companion to Consciousness*, Boston, Mass: Wiley-Blackwell. Aaronson, S., (2014a) “Why I Am Not an Integrated Information Theorist (or, The Unconscious Expander),” *Shtetl Optimized* (blog), May. <https://www.scottaaronson.com/blog/?p=1799>

(2014b) [Giulio Tononi and Me: A Phi-nal Exchange](https://www.scottaaronson.com/blog/?p=1823), *Shtetl Optimized* (blog), June.

Block, N. (1995), “On a Confusion About the Function of Consciousness,” *Behavioral and Brain Sciences* 18, 227–247.

Bostrom, Nick (2104). *Superintelligence: Paths, Dangers and Strategies*. Oxford: Oxford University Press.

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

Harremoes, P., Gee, D. MacGarvin, M., Stirling, A., Keys, J., Wynne, B., and Guedes Vaz, S. (eds.) (2001), *Late lessons from early warnings: the precautionary principle 1896–2000*, Environmental issue report no. 22. Copenhagen: European Environment Agency.

Putnam, H. (1967b). Psychological Predicates. *Art, Philosophy, and Religion.* Pittsburgh, PA, University of Pittsburgh Press.

Epstein 1980, cited in SEP.

Schneider, S. (2017). "It May not Feel Like Anything to be an Alien", Nautilus, Dec.

----(2018). ["Superintelligent AI and the Postbiological Cosmos Approach"](http://schneiderwebsite.com/uploads/8/3/7/5/83756330/9781107175891c09_p178-198.pdf)(2018). Lursch, A.., What is Life? On Earth and Beyond, Cambridge: Cambridge University Press.

----“[How Philosophy of Mind can Shape the Future](http://schneiderwebsite.com/uploads/8/3/7/5/83756330/how_philosophy_of_mind_can_shape_the_fut.pdf),” forthcoming, Philosophy of Mind in the 20th and 21th Century,  Amy Kind (ed.). NY: Routledge.

----(2016). "[Alien Minds](http://www.datascienceassn.org/sites/default/files/Alien%20Minds%20-%20Susan%20Schneider.pdf)." In Dick, S.J. (ed.), The Impact of Discovering Life Beyond Earth, Cambridge University Press, pp. 189–206.

Schneider, S., and Turner, E. (2017), “Is Anyone Home? A Way to Find Out If AI Has Become Self-Aware,” *Scientific American*, July.

Selmer Bringsjord & Paul Bello, “Toward Axiomatizing Consciousness,” (ms.)

Boly, M., Massimini, M., Tsuchiya, N., Postle, B., Koch, C., Tononi, G., (2017) “Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? Clinical and neuroimaging evidence,” Journal of Neuroscience 4 October, 37 (40) 9603-9613.

Hume, D., Treatise of Human Nature, Oxford University Press (1738)

Tononi G, Koch C. 2015 “Consciousness: here, there and everywhere?” *Phil. Trans. R. Soc.* B 370: 20140167.

Tononi, G., Boly, M., Massimini, M. & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience* volume 17, pages 450–461.

More, M.,(1995). *The Diachronic Self: Identity, Continuity, and Transformation*, PhD Dissertation, University of Southern California.

---- "Transhumanism: Towards a Futurist Philosophy". Archived from the original on 29 October 2005.

UNESCO/COMEST, (2005) The Precautionary Principle [The Precautionary Principle](http://unesdoc.unesco.org/images/0013/001395/139578e.pdf) UN: <http://unesdoc.unesco.org/images/0013/001395/139578e.pdf>

Alexander Chislenko, Max More, Anders Sandberg, Natasha Vita-More, Eliezer Yudkowsky, Arjen Kamphius, Nick Bostrom.

“Transhumanist FAQ” <https://humanityplus.org/philosophy/transhumanist-faq/>

Moore, G. E. (1965), "Cramming more components onto integrated circuits", Electronics 38 (8).

Moravec, H. (1989), Mind Children. Harvard: Harvard University Press.

——— (1999), Robot: Mere Machine to Transcendent Mind. New York: Oxford University

Press.

More, M. Principles of Extropy, Version 3.11 2003. <http://www.extropy.org/principles.htm>.

Newman, W. R. (2004), Promethean ambitions: alchemy and the quest to perfect nature. Chicago: University of Chicago Press.

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

Rees, M. (2003), Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century - On Earth and Beyond: Basic Books.

Nietzsche, F. W. (1908), Also sprach Zarathustra: ein Buch fèur alle und keinen. Leipzig: Insel-Verlag.

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

Bostrom, N. (2002), "When Machines Outsmart Humans", Futures 35 (7):759-764.

——— (2003), "Are You Living in a Computer Simulation?" Philosophical Quarterly 53

(211):243-255.

——— (2003), "Human Genetic Enhancements: A Transhumanist Perspective", Journal of

Value Inquiry 37 (4):493-506.

——— The Transhumanist FAQ: v 2.1. World Transhumanist Association 2003.

http://transhumanism.org/index.php/WTA/faq/.

——— (2005), "In Defence of Posthuman Dignity", Bioethics forthcoming.

Bostrom, N. (1998), "How Long Before Superintelligence?" International Journal of Futures Studies 2.

——— (2002), "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards", Journal of Evolution and Technology 9.

Armstrong, D.M. (1989a). Universals: an Opinionated Introduction, Westview Press.

—— (1989b) Nominalism and Realism, vol. 1 of Universals and Scientific Realism, Cambridge University Press.

Chalmers, David J. (2002). “Consciousness and its Place in Nature” in Philosophy of Mind: Classical and Contemporary Readings. Oxford: Oxford Univ. Press.

Churchland, P.(1995). Engine of Reason, Seat of the Soul. Boston: MIT Press.

Clark, A. Natural Born Cyborgs. Oxford: Oxford University Press.

Collins, Nick. (2013). “Hawking: ‘In the future brains could be separated from the body’” The Telegraph. 20 Sep.

Corabi, J & Susan Schneider (2012). “Metaphysics of Uploading.” Journal of Consciousness Studies 19 (7): 26.

Davies, Paul (2010). The Eerie Silence: Renewing Our Search for Alien Intelligence, BostonL Houghton Mifflin Harcourt.

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

Field, H. (1980). Science Without Numbers, Princeton, NJ: Princeton University Press.

Fodor, J. (1998). Concepts: Where cognitive science went wrong. Oxford University Press.

Fodor, J. (2008). LOT 2: The language of thought revisited. Oxford University Press.

Fodor, J. & Brian P. McLaughlin (1990). Connectionism and the Problem of Systematicity. Cognition 35 (2):183-205.

Heil, J. (2005). From an Ontological Point of View. Oxford: Oxford University Press.

Kim, Jaegwon, (2006). Philosophy of Mind, 2nd ed., New York: Westview.

—— (2005). Physicalism, Or Something Near Enough, Princeton University Press.

Laurence, S. & Margolis, E. (2002) Radical concept nativism. Cognition 86, pp. 25-55.

Lewis, D. (1994). Philosophical Papers, vol. 1, New York: Oxford University Press.

Lowe, E.J., (1996). Subjects of Experience. Cambridge University Press.

—— (2006). “Non-Cartesian Substance Dualism and the Problem of Mental Causation,” Erkenntnis, Vol. 65, No. 1, pp. 5-23.

Martin, C. B. (1997). On the need for properties: The road to Pythagoreanism and back. Synthese, 112, 193–231.

Margolis, E. & Laurence, S., eds. (1999) Concepts: Core readings. MIT Press.

McLaughlin, Brian (2007). “Type Materialism about Phenomenal Consciousness” in Velmans and Schneider, The Blackwell Companion to Consciousness, Oxford: Blackwell Publishing.

Prinz, J. (2002) Furnishing the mind: Concepts and their perceptual basis. MIT Press.

Schneider, S. (2007). “What is the Significance of the Intuition that the Laws of Nature Govern?” Australasian Journal of Philosophy, June, pp. 307-324.

—— (2009a). Science Fiction and Philosophy. Oxford: Blackwell Publishing.

—— (2009b). Mindscan: Transcending and enhancing the human brain In (Susan Schneider, ed.), Science Fiction and Philosophy. Wiley-Blackwell. 241-256.

—— (2009c) The nature of symbols in the language of thought. Mind and Language 24(4), pp. 523 – 553.

-------(2011), The Language of Thought: New Philosophical Directions, MIT Press.

------(2012). “Why Property Dualism Cannot Accept Physicalism about Substance.” Philosophical Studies, Vol. 157, Number 1, Jan.

------(2013). “Non-reductive Physicalism and the Mind Problem” (2013a), Nous, Vol. 47, Number 1, pp. 135-153.

------(2015). “Alien Minds,” Discovery (an astrophysics trade anthology, based on a NASA/Library of Congress Symposium), Steven Dick, Cambridge University Press.

-------(2017). “Future Directions for Philosophy of Mind,” in Philosophy of Mind in the 20th and 21th Century, Amy Kind (ed.), forthcoming with Routledge.

—— (forthcoming a.). “How the Mathematical Nature of Physics Undermines Physicalism,” Journal of Consciousness Studies (Target paper for special issue), forthcoming.

—— (forthcoming b.) “Idealism, or Something Near Enough,” in Idealism, Pearce. K and Goldsmidt, T., Oxford: Oxford University Press.

------(forthcoming c),“The Mind is not the Software of the Brain (Even if the Brain is Computational” (ms.)

------(forthcoming d). Superintelligent AI and the Postbiological Cosmos Approach” in Lursch, A. What is Life? Cambridge: Cambridge University Press.

Schneider and Corabi, (2014). “The Metaphysics of Uploading”, Intelligent Machines, Uploaded Minds, Russell Blackford (ed.), Wiley-Blackwell.

Shostak, S. (2009). Confessions of an Alien Hunter, New York: National Geographic.

Smart, J.J.C. (1959). “Sensations and Brain Processes.” Philosophical Review 68 (1959): 141-156.

Strawson, G. (2006). “Realistic monism: Why physicalism entails panpsychism.”Journal of Consciousness Studies. 3 (10---11):3---31.

Turing, A.M. (1950). “Computing Machinery and Intelligence.” Mind 49: 433-460.

Velmans, M. and Schneider, S., (2007). The Blackwell Companion to Consciousness, (with Max Velmans), Oxford: Blackwell Publishing.

1. Cite the film *Contact*, 1997. [↑](#endnote-ref-1)
2. Bostrom, N. *Superintelligence: Paths,Dangers, Strategies.* [↑](#endnote-ref-2)
3. Cite article at https://www.technologyreview.com/s/611077/the-worlds-most-powerful-supercomputer-is-tailor-made-for-the-ai-era/ [↑](#endnote-ref-3)
4. **Schneider, Susan. “It may not feel like anything to be an alien.” *Nautilus*, Dec. 2016.**  [↑](#endnote-ref-4)
5. More, M., (1990), In a 1990 essay "Transhumanism: Toward a Futurist Philosophy. Bostrom, Nick (April 2005). "A history of transhumanist thought" (PDF). *Journal of Evolution and Technology*. 14 (1): 1–25. [↑](#endnote-ref-5)
6. Kurzweil, R., *The Singularity is Near*. [↑](#endnote-ref-6)
7. The problem of AI consciousness is also distinct from a philosophical problem called “the problem of other minds.” Notice that each of us can tell, by introspecting, that we are conscious, but we cannot tell, from introspection, that the other humans around us are conscious beings. This is a well-known form of philosophical skepticism. We tend to infer that other normal humans are conscious, because they are made of they have nervous systems like our own, and they exhibit the same kinds of behaviors, such as wincing when in pain, seeking friendships, and so on. AI, in contrast, will not have nervous systems like our own, and they may behave in quite alien ways. [↑](#endnote-ref-7)
8. “Biological naturalism,” as used here, doesn’t involve Searle’s larger position about physicalism and the metaphysics of mind. But for a discussion see his piece in Schneider and Velmans, X. “Biological Naturalism.” For our purposes, Biological Naturalism is just a generic position saying that machines cannot be consciousness, where “cannot” refers to conceptual possibility, or, in a weaker version of Biological Naturalism, nomological possibility. It is worth noting that Searle himself seemed sympathetic to the possibility of neuromorphic computation being conscious (Schneider and Velmans, X). [↑](#endnote-ref-8)
9. **For an influential depiction of techno-optimism see Ray Kurzweil, *The Age of Spiritual Machines*, and *The Singularity is Near*.**  [↑](#endnote-ref-9)
10. A distinctive element of computational explanations in cognitive science is that cognitive and perceptual capacities are understood by decomposing them into more basic operations, which are themselves decomposable into more basic constituents, which themselves can be explained causally (Block 1995). This leading explanatory approach in cognitive science has been called the *method of functional decomposition* because it explains the properties of a system by decomposing it into the causal interaction between constituent parts, which are themselves often explained by the causal interaction between their own subsystems. [↑](#endnote-ref-10)
11. The following chapter will discuss, in more detail, what is meant by “same computational configuration.” It will turn out that there are different levels of precision that can be employed. [↑](#endnote-ref-11)
12. In philosophical jargon, such would be called a “precise functional isomorph.” [↑](#endnote-ref-12)
13. I’m simplifying things by merely discussing neural replacement in the brain. For instance, perhaps neurons elsewhere in the nervous system, such as the gut, are relevant as well. This kind of thought experiment could be modified to suppose that more than neurons in the brain are replaced. [↑](#endnote-ref-13)
14. Chalmers, David, *The Conscious Mind*. Oxford: Oxford University Press. (cite Lowe and Plantinga) [↑](#endnote-ref-14)
15. Here, I am assuming that biochemical properties could be included. In principle, if they are relevant to cognition, an abstract characterization of the behavior of such features could be included in a functional characterization. [↑](#endnote-ref-15)
16. Get FDA forms from DR. [↑](#endnote-ref-16)
17. A complete, precise copy could occur in the context of brain uploading, however. Like the case of the isomorph of you, human brain uploading remains far in the future. [↑](#endnote-ref-17)
18. Kurzweil, *The Singularity is Near*, *p.* 375. See also 361, 362, 364, 387, and 476. [↑](#endnote-ref-18)
19. Feb. 17, 2018. [↑](#endnote-ref-19)
20. For an intriguing case for ethical obligations to conscious AIs see “A Defense of the Rights of Artificial Intelligences,” Eric Schwitzgebel and Mara Garza, Midwest Studies in Philosophy, Sept. 2015. [↑](#endnote-ref-20)
21. https://www.wired.com/2016/12/neuroscientist-whos-building-better-memory-humans/ [↑](#endnote-ref-21)
22. Murray Shanahan, “Conscious Exotica,” Aeon, https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there [↑](#endnote-ref-22)
23. Peter Singer, Animal liberation: a New Ethics for the Treatment of Animals” 1975. NY: Harper Collins [↑](#endnote-ref-23)
24. Block. Axioms for FC in highly intelligent AI have been formulated by Selmer Brinsjord (cite). [↑](#endnote-ref-24)
25. Crick and Koch, 1990. https://www.biorxiv.org/content/biorxiv/early/2017/03/19/118273.full.pdf [↑](#endnote-ref-25)
26. Cite these. https://www.tnu.ethz.ch/fileadmin/user\_upload/Future\_directions.pdf. https://www.biorxiv.org/content/biorxiv/early/2017/03/19/118273.full.pdf [↑](#endnote-ref-26)
27. <http://www.nytimes.com/2010/09/21/science/21consciousness.html?pagewanted=all> [↑](#endnote-ref-27)
28. Tononi and Koch, 2015, “Consciousness: here, there and everywhere? Cite also the piece here: https://www.nature.com/articles/nrn.2016.44. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0

    PLoS Computational Biology (2014) [↑](#endnote-ref-28)
29. I will subsequently refer to this level of phi rather vaguely, as “high Φ” because calculations of phi for the biological brain are currently intractable. [↑](#endnote-ref-29)
30. [↑](#endnote-ref-30)
31. Readers familiar with Chalmers “absent qualia, dancing qualia” thought experiment may object that I’ve missed something, for Chalmers’ thought experiment supports the view that consciousness supervenes on functional configuration: if you fix the psychofunctional facts, you fix the qualia. But I am disputing that functional isomorphism occurs in the first place. I 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. [↑](#endnote-ref-31)
32. It should be noted that transhumanism by no means endorses every sort of enhancement. For example, Nick Bostrom rejects positional enhancements (enhancements primarily employed to increase one’s social position) yet argues for enhancements that could allow humans to develop ways of exploring “the larger space of possible modes of being” (2005: 11). [↑](#endnote-ref-32)
33. Kurzweil 1999, 2005; Bostrom 2003, 2005 [↑](#endnote-ref-33)
34. There are many nuances to this rough trajectory. For instance, some transhumanists believe that the move from unenhanced human intelligence to superintelligence will be extremely rapid because we are approaching a singularity, a point at which the creation of superhuman intelligence will result in massive changes in a very short period (e.g. 30 years) (Bostrom 1998; Kurzweil 1999, 2005; Vinge 1993). Other transhumanists hold that technological changes will not be so sudden. These discussions often debate the reliability of Moore’s Law (Moore 1965). Another key issue is whether a transition to superintelligence will really occur because the upcoming technological developments involve grave risk. The risks of biotechnology and AI concern transhumanists, progressive bioethicists more generally, as well as bioconservatives (Annis 2000; Bostrom 2002; Garreau 2005; Joy 2000). [↑](#endnote-ref-34)
35. For mainstream anti‐enhancement positions on this question see, e.g., Fukuyama (2002), Kass et al. (2003), and Annas (2000). [↑](#endnote-ref-35)
36. Kurzweil 2005: 383. [↑](#endnote-ref-36)
37. There are different versions of psychological continuity theory. One could, for instance, appeal to (a): the idea that memories are essential to a person. Alternatively, one could adopt (b), one’s overall psychological configuration is essential, including one’s memories. Herein, I shall work with one version of this latter conception – one that is inspired by cognitive science – although many of the criticisms of this view will apply to (a) and other versions of (b) as well. For some different versions see chapter 27 of John Locke’s 1694 *Essay Concerning Human Understanding* (note that this chapter first appears in the second edition; it is also reprinted as “Of Identity and Diversity” in Perry 1975). See also the essays by Anthony Quinton and Paul Grice, both of which are reprinted in Perry (1975). [↑](#endnote-ref-37)
38. Kurzweil 2005: 383 [↑](#endnote-ref-38)
39. Brain-based Materialism, as discussed here, is more restrictive than physicalism in philosophy of mind, for a certain kind of physicalist could hold that you could survive radical changes in your substrate, being brain-based at one time, and becoming an upload at a later time. [↑](#endnote-ref-39)
40. Sociologist James Hughes holds a transhumanist version of the no self view. (See the Institute for Ethics and Emerging Technology’s “Cyborg Buddha” project at <http://ieet.org/index.php/IEET/cyborgbuddha.)> For helpful surveys of these four positions, see Eric Olson’s chapter in my volume, Science Fiction and philosophy (Chapter 7), and Conee and Sider (2005)*.* [↑](#endnote-ref-40)
41. Meditations in First Philosophy, Descartes [↑](#endnote-ref-41)
42. This is a version of a computational theory of mind which I will criticize in Chapter Six. Further, computational theories of mind can appeal to various computational theories of the format of thought: connectionism, dynamical systems theory (in its computational guise), the symbolic or language of thought approach, or some combination thereof. These differences will not matter to our discussion. [↑](#endnote-ref-42)
43. Kurzweil 2005: 383 [↑](#endnote-ref-43)
44. Cite Hughes paper in Transhumanist Reader and his Cyborg Buddha book [↑](#endnote-ref-44)
45. Chapter Eight will discuss the transhumanists’ computational approach to the mind in more detail. Many advocates of computational theories of the mind do not develop it as a theory of personal identity. An advocate of CTM could remain silent about the topic of personal identity altogether. But, in practice, computationalism is the underlying perspective on the mind that helps motivate the transhumanist’s patternism. [↑](#endnote-ref-45)
46. Sawyer 2005: 44–5 [↑](#endnote-ref-46)
47. Sawyer 2005: 18. [↑](#endnote-ref-47)
48. **Here I am indebted to the groundbreaking work by Paul Davies, Steven Dick, Martin Rees, and Seth Shostak, among others.** [↑](#endnote-ref-48)
49. Shostak, 2009; Davies, 2010; Dick, 2013; Schneider, 2015. [↑](#endnote-ref-49)
50. http://science.sciencemag.org/content/131/3414/1667 [↑](#endnote-ref-50)
51. Cite Schneider, “Alien Minds” in Dick, S., *Discovery*, Cambridge: Cambridge Univ. Press. [↑](#endnote-ref-51)
52. Cite. https://www.theguardian.com/science/2013/sep/21/stephen-hawking-brain-outside-body [↑](#endnote-ref-52)
53. Harmon, Amy, NYT cite: <https://www.nytimes.com/2015/09/13/us/cancer-immortality-cryogenics.html> [↑](#endnote-ref-53)
54. See

    <https://www.nytimes.com/2015/09/13/us/cancer-immortality-cryogenics.html>

    https://alcor.org/Library/html/casesummary2643.html [↑](#endnote-ref-54)
55. Cite <https://www.technologyreview.com/s/542601/the-science-surrounding-cryonics/> [↑](#endnote-ref-55)
56. Harmon (2015). [↑](#endnote-ref-56)
57. I’m grateful to Kim Souzzi’s boyfriend, Josh Schisler, for a helpful email conversation on this matter (Aug. 26, 2018). See also Harmon, (2015). [↑](#endnote-ref-57)
58. Harmon 2015 [↑](#endnote-ref-58)
59. Harmon, (2015). [↑](#endnote-ref-59)
60. Schneider 2014, Corabi and Schneider 2014. For an overview of different steps of uploading see the piece accompanying the article on Souzzi. “The Neuroscience of Immortality.” https://www.nytimes.com/interactive/2015/09/03/us/13immortality-explainer.html?mtrref=www.nytimes.com&gwh=38E76FFD21912ECB72F147666E2ECDA2&gwt=pay [↑](#endnote-ref-60)
61. Schneider 2014; Corabi and Schneider 2014. We see similar behavior in the quantum world, but we have no reason to believe that ordinary mid-sized physical objects, such as tables, trees and brains, exhibit such behavior. [↑](#endnote-ref-61)
62. I will focus on claims by fusion-optimists, as they are the ones making claims about radical enhancement. Many academics appealing to the Software View are more interested in characterizing how the mind works, rather venturing claims about brain enhancement or uploading. [↑](#endnote-ref-62)
63. Wiley 2014 [↑](#endnote-ref-63)
64. http://www.brainpreservation.org/ken-hayworths-personal-response-to-mit-technology-review-article/ [↑](#endnote-ref-64)
65. The notion of an implementation has been problematic. See Putnam 1967, Piccinini (plug citation in from http://www.umsl.edu/~piccininig/Computational\_Functionalism.htm) [↑](#endnote-ref-65)
66. Block 1995, 140. [↑](#endnote-ref-66)
67. [↑](#endnote-ref-67)