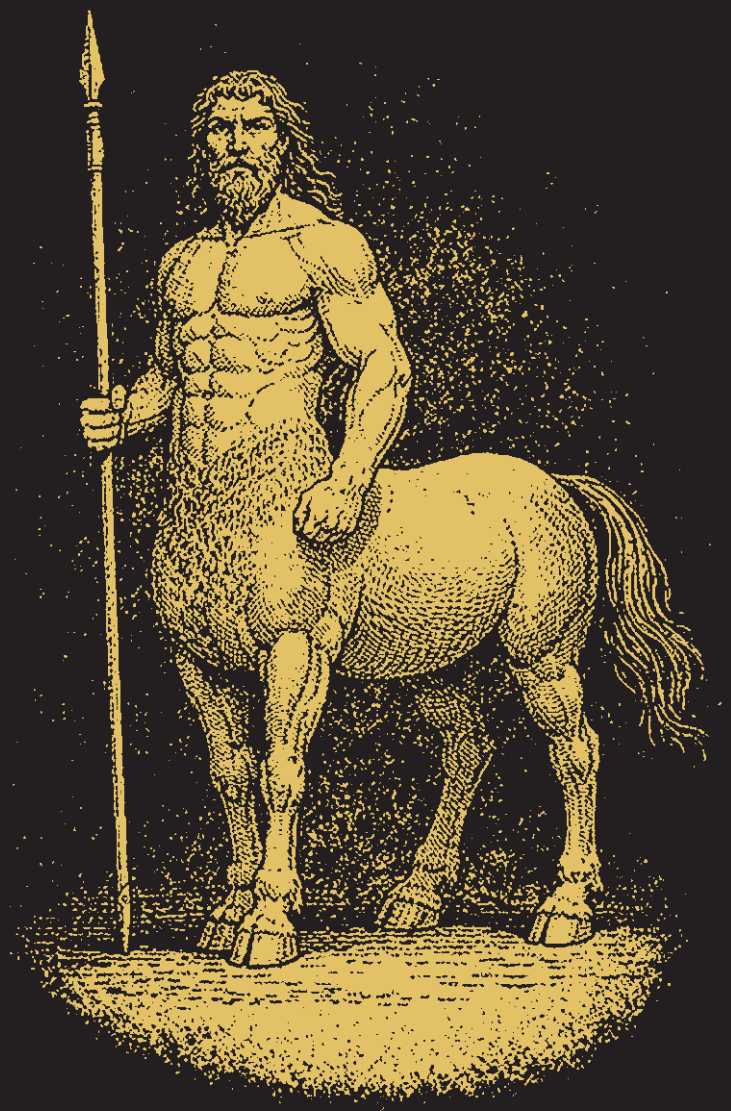

Centaurus

3

Susan Schneider

The future of the mind

A thought experiment

It's 2045, and you're out shopping. Your first stop is the Center for Mind Design. As you walk in, a large menu stands before you, listing brain enhancements with funky names. Hivemind is a brain chip that allows you to experience the innermost thoughts of your loved ones. Zen Garden is a microchip that induces zen-master-level meditative states. Human Calculator is an implant that gives you savant-level mathematical abilities. What would you select, if anything? Enhanced attention? Mozart-level musical skills? You can order a single enhancement or bundle several together.

So that's the scenario. As a philosopher, I like thought experiments because they really get our imagination going. This text is about the future of your mind, and how our understanding of ourselves, our minds, and our nature can change the future for better or worse. What I'll be asking is, at a place like this hypothetical Center for Mind Design, could you truly merge with artificial intelligence?

This text explores key issues surrounding transhumanist ideas, neurotechnology, and the age-old philosophical questions they invoke. It begins with an examination of the foundational concepts, providing the necessary context for understanding the broader implications. From there, the discussion shifts to the potential limits of human intelligence augmentation, particularly in relation to consciousness and personhood. Finally, the chapter concludes with a reflection on these advancements within the larger framework of our place in the cosmos.

Mind design

Let's turn to some quick background. Brain evolution was heavily constrained by environmental, anatomical, and metabolic demands, but AI-based brain enhancement technologies could augment intelligence at a rate that's much faster than biological evolution. I call this new enterprise "mind design".

It's a kind of intelligent design, but we are not some Gods pur-

porting to be the designers. Our social development often seems to lag behind our technological capacities.

Now suppose you're standing there at this Center for Mind Design. You're waiting in line, and somebody comes up to you, and they say: "You know what? You should come back in a few weeks because if clinical trials go as planned, customers will soon be able to purchase an enhancement bundle called MERGE, a series of enhancements allowing you to gradually augment and transfer all of your mental functions to the cloud over a period of five years". So could you do this? Could you really begin to remove parts of your body? Could you really begin to remove parts of your biological brain and transfer your mental capacities into the cloud, the Internet, or into another computer? And would that be you? Would you survive?

So what does it mean to merge with artificial intelligence in the first place and why should we worry about it today?

Transhumanists believe that humans can overcome their biological limitations through science and technology. The idea itself shouldn't seem objectionable to most of us. However, the devil's in the details. The transhumanist trajectory for enhancement is much more controversial. So they start with this idea of an unenhanced human. The unenhanced human has significant upgrading with cognitive and other physical enhancements.

So perhaps you get the ability to have Mozart or Philip Glass level musical composition, or you get the ability to expand your attention so that you can focus on hundreds of things. After replacing so much of the biological brain with brain chips, you become superintelligent AI. You're no longer technically human. You're posthuman.

The idea of neural replacement

A superintelligent AI is a hypothetical form of AI that outthinks humans in every way possible: scientific reasoning, social intelligence, and more. So this kind of trajectory is the kind of view that's held by a lot of transhumanists such as Nick Bostrom, Elon Musk, Ray Kurzweil, who's now a chief engineer at Google, and the late Stephen Hawking.

I want to focus on one aspect of transhumanism that I find to be most problematic. And that is the idea that augmenting or enhancing intelligence involves replacing parts of the biological brain with AI components—the idea of neural replacement. That includes brain uploading or a gradual process in which parts of your brain are outmoded and transferred over to computer chips or computers.

Transhumanism is a very influential view, and it connects in a really interesting way with the classic philosophical issue—does consciousness transcend the brain? There is a similar view, in cognitive science where it says that the mind is something like a program. But I consider this idea flawed and believe we should think of these issues to not commit what the historian Michael Best called the Jetsons fallacy.

You may have seen *Star War*s or the cartoon The Jetsons and noticed in these science fiction stories, humans are depicted as not being enhanced, but they're surrounded by all of these robots and self-driving cars. But if you really think about what the future is like, it would be sort of surprising if all kinds of artificial intelligence is all around us. We have sophisticated robots taking care of our homes and autonomous vehicles driving us places, but we don't use that very same technology to augment our own intelligence.

In fact, that's already what's going on, with the military. There are various programs in the United States to create super soldiers.

Elon Musk established a company called Neuralink. He hopes that humans can keep up with artificial intelligence by having some sort of merger of biological intelligence and machine intelligence. And to this end, he created Neuralink, which already has

implantable chips allowing data to travel wirelessly from the brain to digital devices.

There is also a project that is very far along in clinical trials in the United States to build an artificial hippocampus. This is Ted Berger's fascinating project. The hippocampus is the part of the brain responsible for forming new memories and is crucial to memory formation. Facebook is working on this, as are Google and Kernel. In fact, the list of companies involved is quite long.

What does all this mean for the human future?

Let's think about these issues and the future of the self and mind from a philosophical standpoint. The transhumanist view on this topic is that we should engage in these kinds of experiments, and people should be free to augment their intelligence at a place like the Center for Mind Design. One reason is that it is exciting to imagine being free from unwanted psychological elements: to sculpt our own character, to remove chronic depression, and to make deliberate decisions about our psychology. Second, this offers another route to longevity and a possible path toward intelligence augmentation and merging with artificial intelligence.

But on the other hand, it doesn't take too long for us to feel deeply concerned that our thoughts could be part of a computer system connected to the Internet, given what's happening with social media companies using our data to manipulate us and selling our private information. It could be a complete dystopia—a thought data economy. There are many highly negative examples, such as Facebook, which utilizes people's personal information and employs social psychology techniques to manipulate users and foster addiction to their platforms, without regard for the impact on different countries and groups.

A surveillance economy that possesses neural data is highly dangerous. It is also a serious threat in non-capitalist systems, such as authoritarian dictatorships like China, where the technology could be used to monitor and control dissidents. Now, let's turn to the philosophical issues beyond ethics. Even if we were able to establish privacy laws and regulations, there are still reasons to believe that the transhumanist view of merging with AI may be unattainable.

Design ceilings

There are limits to human intelligence enhancement that are not imposed by evolution, medicine, or technology. There are also philosophical limits—perhaps one reason why those developing brain chips haven't fully considered the implications. Of course, there will be other limits as well, such as neurotechnological and medical ones. But what kind of limits do I have in mind?

The first limit is what I refer to as the 'consciousness ceiling'—a boundary that emerges if microchips prove incapable of supporting conscious experience. The second, the 'self ceiling,' marks the point at which an individual seeking enhancement undergoes such profound transformation that they cease to exist as the person they once were. In this case, the procedure does not simply modify the mind; it fundamentally alters the identity of the one who undertakes it.

Consciousness is the felt quality of experience. So when you see the rich use of a sunset or you smell the aroma of your morning coffee or you stub your toe, it always feels like something to you from the inside. And nobody else knows exactly what it feels like to be you. You can try to communicate it, but it's really a private thing inside of your head. It's what it feels like from the inside to be you. You're conscious all throughout your waking life and even when you're dreaming.

Richard Dawkins and I appeared together in a film called *Super Sapiens*, in which he made a particularly provocative remark: "It's not obvious to me that a replacement of our species by our own technological creations would necessarily be a bad thing."

You might find this statement open-minded—and in a way, it is. It's not necessarily a bad thing. I often hear people make this claim, suggesting that humanity is following an evolutionary trajectory in which we eventually give way to something superior. Perhaps a being with greater forms of conscious experience, one that transcends our flaws and surpasses us in ways we cannot yet imagine.

This could be the direction of the future, but let's think about it in greater detail. I wouldn't encourage you to think about whether machines can be conscious. Because if nonconscious machines supplanted biological intelligence, then this singularity, this idea of a technological artificial intelligence explosion that makes our lives better is actually a nightmare. It wouldn't be a transhumanist utopia the way people like Elon Musk and Ray Kurzweil discussed. Instead, it would be the end of consciousness on earth. And this issue is not explored.

Let's go back to the Center for Mind Design and remember the thought experiment. Suppose you have the opportunity to pay thousands of dollars to purchase an enhancement bundle called MERGE that allows you to gradually replace parts of your brain with microchips, eventually leading to your being uploaded to the cloud or to your brain being entirely replaced by microchips.

If AI is not conscious, then merging with it would be a bad idea, because if microchips are the wrong substrate for consciousness, a mind-machine merger would not preserve it. If you tried, you would lose your consciousness, and your mind would cease to exist. That would represent a clear limit to human intelligence augmentation.

If this is true, we will not be able to surpass or even match the intelligence of artificial intelligence in the future. As AI becomes increasingly advanced—potentially reaching and exceeding human-level intelligence—we will be unable to keep up if a consciousness ceiling exists.

Now, let's consider another potential limitation on intelligence enhancement, one even more significant than the consciousness

ceiling. While it may be possible to develop microchips that support consciousness, the concept of self-ceiling is beyond the reach of scientific solutions. It is a purely philosophical issue.

For the sake of discussion, let's assume that the consciousness ceiling does not arise. Should you proceed and invest further?

To determine whether you should enhance yourself, you must first understand what you are to begin with. But what does it mean to be a person or a self? Would you continue to exist if parts of your brain were replaced with microchips, or would you unknowingly end your own existence, only to be replaced by someone—or something—else?

To grasp the depth of this problem, it is helpful to explore the literature in contemporary philosophy. These questions trace back to thinkers like Locke, Hume, and Nietzsche.

Recent philosophical thought explores the metaphysics of everyday objects. Consider an espresso machine: if it's unplugged, it's still the same machine. But suppose I use a futuristic, science-fiction-style gun to disintegrate it into dust. Is it still the same machine? You would likely say no—it no longer exists. And now, you'll never get your coffee.

Notice that certain features of the coffee machine are essential for it to continue existing. The same applies to us. Suppose you are religious and believe in a soul—then the soul would be essential to your continued existence. Or, if you believe your existence depends on your brain and nervous system, then if your brain were destroyed, you would recognize that having a brain is essential to your survival. Without it, you would cease to exist.

This simple yet crucial observation has often been overlooked in discussions about the future of the mind.

If brain enhancement makes a person super-intelligent, that may seem exciting. But it cannot come at the cost of eliminating any features essential to your survival. Otherwise, you might walk

into a mind-design center—but you won't walk out. You will have unknowingly ended your own existence.

You cannot truly merge with artificial intelligence. So even if the technology works, you would be paying for a smarter mind or a fitter body—but it wouldn't be you. It would be something else, perhaps a digital twin. But it wouldn't be you. You would be gone. That would be a disturbingly perverse direction for technology to take. Medicine is meant to help people, and such an outcome would be horrifying.

I call this the self ceiling. It marks the point beyond which a person who seeks enhancement is no longer the same individual, as the procedure causes the original self to cease to exist.

The nature of the self

It's a philosophical question about what the nature of the self is. It's a question that's been debated since the beginning of philosophical thinking. And we probably all have a sense that it's very difficult to prove any one position on the nature of the self, soul, or mind because it's very controversial, and it's not like you can run an experiment to find out if there's a God or to find out if the mind is a program.

One way to describe the issue is to indicate that there are lots of different theories out there on the nature of the self or person. There's materialism—the idea that you are essentially your brain and nervous system. There's what's called the psychological continuity view, which is a view held by the philosopher Locke that says that you are your memories and your ability to reflect on yourself and your overall psychological configuration. The transhumanists have a version of that view that they tend to endorse. That is one called patternism.

Patternism is a view that was articulated by Ray Kurzweil and also Nick Bostrom in their work. It says that what is essential to you is your computational configuration, the sensory systems

that your brain has, the association areas that are responsible for integrating the different subsystems of the brain, the neural circuitry making up your reasoning abilities, your memories, and so on. Together, these form a sort of algorithm that describes how your brain computes, and it's unique to you. It's your pattern.

There is also a view that we have souls, and obviously many religions hold this position. And there's also a religious and philosophical view that says that the self is an illusion, that there's no underlying self there, and there's also no surviving person from moment to moment.

The important thing to note is that each of these views has an answer. For example, For example, the materialist view would say you'd be killing yourself and you shouldn't do it. Instead of developing brain chips that replace parts of the biological brain or trying to upload your intelligence, if you want to enhance, you should take a different path forward. Instead, the view here would be that neurotechnology should develop biological brain enhancements and minimal AI enhancements that don't replace or damage key parts of the brain.

This is not a conception of the future in which we merge with artificial intelligence. If AI ultimately outthinks us, we would simply be unable to keep up. If you have a soul theory, on the other hand, it's not at all clear whether you should enhance. That would depend on the details of the religion, so you would have to have discussions with your religious adviser. Well, what if you have no self view? You never merge with AI because there's no you. But you could strive to enhance that view. Because if you truly believe you don't survive, you'd be open to the possibility of replacing your brain with microchips. The trouble is knowing which view is right, having the certainty required to risk your life.

What about patternism? Here's the problem. When does the pattern begin, and when does the pattern end? Maybe deleting some bad chess-playing habits or uploading AlphaGo is okay?

Maybe you can get away with that? You wouldn't change your whole nature. But what if you bought merch, or what if you purchased several different enhancements? At what point would you no longer continue to be you? And how could we find out scientifically? How could we find out with certainty?

This is a philosophical issue, and I doubt that the science of brain chips will resolve it. That's why I adopt a stance of metaphysical humility. Claims involving mind transfer to a new substrate—such as a computer—or drastic alterations to the brain must be scrutinized carefully. Bear in mind that there is intense debate within the philosophical literature.

The future of intelligence

Suppose alien intelligence exists. Based on astrophysical projections, Earth is a relatively young planet, meaning an alien civilization could be 50,000 years more advanced than us. They would have already developed artificial intelligence and augmented their own intelligence. Planets across the universe may have already grappled with the same questions I raised earlier. When considering the future of intelligence, we should ask: Could intelligent aliens actually be forms of artificial intelligence? And if so, are they even conscious?

It ultimately depends on the mind-design choices that a culture makes. If design ceilings exist, then the issues I raised earlier function like philosophical laws—constraints on biological systems that may prevent them from augmenting in ways that keep pace with artificial intelligence.

This presents a different perspective on evolution. When discussing the future of intelligence, we are no longer in the realm of Darwinian evolution but rather in that of intelligent—or even unintelligent—design. In this new realm, humans and other biological beings attempt to augment intelligence, but such evolution will still face constraints. These include design ceilings, economic

limitations affecting tech companies developing microchips and AI systems, and regulatory constraints imposed by AI laws.