ARTIFICIAL YOU

AI AND THE FUTURE OF YOUR MIND

SUSAN SCHNEIDER

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

—-1
—0
—+1

CONTENTS

—-1
—0
—+1

INTRODUCTION

It is 2045. Today, you are out 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. "Hive Mind" is a brain chip allowing you to experience the innermost thoughts of your loved ones. "Zen Garden" is a microchip for Zen master-level meditative states. "Human Calculator" gives you savant-level mathematical abilities. What would you select, if anything? Enhanced attention? Mozart-level musical abilities? You can order a single enhancement, or a bundle of several.

Later, you visit the android shop. It is time to buy that new android to take care of the house. The menu of AI minds is vast and varied. Some AIs have heightened perceptual skills or senses we humans lack, others have databases that span the entire Internet. 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. Our brains evolved for specific environments and are greatly constrained by anatomy and evolution. But artificial intelligence (AI) has opened up a vast design space, offering new materials and modes of operation, as well as novel 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 of intelligent design, but we humans, not God, are the designers.

—-1
—0
—+1

I find the prospect of mind design humbling, because frankly, we are not terribly 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] We walk the moon, we harness the energy of the atom, yet racism, greed, and violence are still commonplace. Our social development lags behind our technological prowess.

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 also a cost to not understanding issues in philosophy, 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 mental life. Without consciousness, there would be no pain or suffering, no 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 don't have 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 artificial intelligence could lead to the failure of conscious beings to flourish. For if we are not careful, we may experience one or more *perverse realizations* of AI technology—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 already discussed AI-based threats to human flourishing. The threats range from hackers shutting down 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. The perverse realizations I have in mind generally fall into one of the following types: (1) overlooked situations involving the creation of conscious machines and (2) scenarios that concern radical brain enhancement, such as the enhancements at the hypothetical Center for Mind Design. Let's consider each kind of scenario in turn.

CONSCIOUS MACHINES?

Suppose that we create sophisticated, general-purpose AIs: AIs that can flexibly move from one kind of intellectual task to the next and can even rival humans in their capacity to reason. Would we, in essence, be creating *conscious* machines—machines that are both selves and subjects of experience?

When it comes to how or whether we could create machine consciousness, we are in the dark. One thing is clear, however: The question of whether AIs could have experience will be key to how we value their existence. Consciousness is the philosophical cornerstone of our moral systems, being central to our judgment of whether someone or something is a self or person rather than a mere automaton. And if an AI is a conscious being,

forcing it to serve us would be akin to slavery. After all, would you really be comfortable giving that android shop your business if the items on the menu were conscious beings—beings with mental abilities rivaling, or even exceeding, those of an unenhanced human?

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 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 by-product of building an intelligent system, or it may be altogether impossible.

In the long term, the tables may turn on humans, and the problem may not be what we could do to harm AIs, but what AI might do to harm us. Indeed, some suspect that synthetic intelligence 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 to AI, a rung on the evolutionary ladder. For instance, Stephen Hawking, Nick Bostrom, Elon Musk, Max Tegmark, Bill Gates, and many others have raised "the control problem," the problem of how humans can control their own AI creations, if the AIs outsmart us.[2] Suppose we create an AI that has human-level intelligence. With self-improvement algorithms, and with rapid computations, it could quickly discover ways to become vastly smarter than

-1—
0—
+1—

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. This is only one way that synthetic beings could supplant organic intelligences; alternatively, humans may merge with AI through cumulatively significant brain enhancements.

The control problem has made world news, fueled by Nick Bostrom's recent bestseller: *Superintelligence: Paths, Dangers and Strategies.*[3] What is missed, however, is that consciousness could be central to how AI values *us.* Using its own subjective experience as a springboard, superintelligent AI 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. If superintelligent machines are not conscious, either because it's impossible or because they aren't designed to be, we could be in trouble.

It is important to put these issues into an even larger, universe-wide context. In my two-year NASA project, I suggested that a similar phenomenon could be happening on other planets as well; elsewhere in the universe, other species may be outmoded by synthetic intelligences. 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. And should these AIs be incapable of consciousness, as they replace biological intelligences, the universe would be emptied of these populations of conscious beings.

If AI consciousness is as significant as I claim, we'd better know if it can be built, and if we Earthlings have built it. In the coming chapters, I will explore ways to determine if synthetic

—-1
—0
—+1

consciousness exists, outlining tests I've developed at the Institute for Advanced Study in Princeton.

Now let's consider the suggestion that humans should merge with AI. Suppose that you are at the Center for Mind Design. What brain enhancements would you order from the menu, if anything? You are probably already getting a sense that mind design decisions are no simple matter.

COULD YOU MERGE WITH AI?

I wouldn't be surprised if you find the idea of augmenting your brain with microchips wholly unnerving, as I do. As I write this introduction, programs on my smartphone are probably tracking my location, listening to my voice, recording the content of my web searches, and selling this information to advertisers. I think I've turned these features off, but the companies building these apps make the process so opaque that I can't be sure. If AI companies cannot even respect our privacy 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 suppose that AI regulations improve, and our brains could be protected from hackers and corporate greed. Perhaps you will then 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—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."[4] 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 massive computing power is available.

Musk's motivations may be less than purely 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 AI researchers themselves need a better idea of what is at stake.

For instance, if AI cannot be conscious, then if you substituted a microchip for the parts of the brain responsible for consciousness, you would end your life as a conscious being. You'd become what philosophers call a "zombie"—a nonconscious simulacrum of your earlier self. Further, even if microchips could replace parts of the brain responsible for consciousness without zombifying you, radical enhancement is still a major risk. After too many changes, the person who remains may not even be you. Each human who enhances may, unbeknownst to them, 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. They tend to sympathize with a conception of the mind that says the mind is a software program. According to them, you can enhance your brain hardware in radical ways and still run the

same program, so your mind still exists. Just as you can upload and download a computer file, so your mind, as a program, could even be uploaded to the cloud. This is a technophile's route to immortality—the mind's new "afterlife," if you will, that outlives the body. As alluring as a technological form of immortality may be, though, 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 deep 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 be developed. Why suspect any of this will happen?