# ARTIFICIAL YOU

# ARTIFICIAL YOU

## AI AND THE FUTURE OF YOUR MIND

SUSAN SCHNEIDER

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Published by Princeton University Press
41 William Street, Princeton, New Jersey 08540
6 Oxford Street, Woodstock, Oxfordshire OX20 1TR

press.princeton.edu



ISBN 9780691180144
ISBN (ebook): 9780691197777

British Library Cataloging-in-Publication Data is available

Editorial: Matt Rohal
Production Editorial: Terri O'Prey
Text Design: Leslie Flis
Production: Merli Guerra
Publicity: Sara Henning-Stout, Katie Lewis
Copyeditor: Cyd Westmoreland

Jacket design by Faceout Studio, Spencer Fuller
Jacket background: Wizemark / Stocksy

This book has been composed in Arno Pro and Trade Gothic LT Std

Printed on acid-free paper. ∞

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

# CONTENTS

# ARTIFICIAL YOU

# 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.

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

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

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, your mind, as a program, could 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?

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, integrating ideas on various topics, and even, perhaps, outthinking humans. 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 constructing smart household assistants, robot supersoldiers, 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.

Given the current rapid-fire pace of its development, AI may advance to artificial general intelligence (AGI) within the next several 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 decades. According to a recent 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 percent probability by 2050, and within a 90 percent probability by 2070.[1]

In the long run, there is simply no contest. AI will be far more capable and durable than we are.

## THE JETSONS FALLACY

None of this necessarily means that we humans will lose control of AI and doom ourselves to extinction, as some say. If we enhance our intelligence with AI technologies, perhaps we can keep abreast of it. Remember, AI will not just make for better robots and supercomputers. In the film *Star Wars* and the cartoon *The Jetsons*, humans are surrounded by sophisticated AIs, while themselves remaining unenhanced. The historian Michael Bess has called this *The Jetsons Fallacy.*[3] 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. To the contrary, it is a plausible extrapolation of present technological trends.

Increasingly, the human brain is being regarded as something that can be hacked, like a computer. In the United States alone, there are already many projects developing brain-implant technologies to treat mental illness, motion-based impairments, strokes, dementia, autism, and more.[4] The medical treatments of today will inevitably 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 become a cyborg.

## 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."[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. They will have radically advanced intelligence, near immortality, deep friendships with AI creatures, and elective body characteristics. Transhumanists share the belief that such an outcome is very desirable, both for one's own personal development and for the development of our species as a whole. (To further acquaint the reader with transhumanism, I've included the Transhumanist Declaration in the Appendix.)

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 observers in the relevant scientific fields as being on their way.[6] For instance, Oxford University's Future of Humanity Institute—a major transhumanist group—released a report on the technological requirements for uploading a mind to a machine.[7] A U.S. Defense Department agency has funded a

program, Synapse, that is trying to develop a computer that resembles the brain in form and function.[8] Ray Kurzweil has even discussed the potential advantages of forming friendships, *Her*-style, with personalized AI systems.[9] All around us, researchers are striving to turn science fiction into science fact.

You may be surprised to learn that I consider myself a transhumanist, but I do. I first learned of transhumanism while an undergraduate at the University of California at Berkeley, when I joined the Extropians, an early transhumanist group. After poring through my boyfriend's science fiction collection and reading the Extopian listserv, 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.

## A FEW WORDS OF WARNING

The challenge is how to get there from here in the face of radical uncertainty. No book written today could accurately predict the contours of mind-design space, and the underlying philosophical mysteries may not diminish as our scientific knowledge and technological prowess increase.

It pays to keep in mind two important ways in which the future is opaque. First, there are known unknowns. We cannot be certain when the use of quantum computing will be commonplace, for instance. We cannot tell whether and how certain AI-based technologies will be regulated, or whether existing AI safety measures will be effective. Nor are there easy, uncontroversial answers to the philosophical questions that we'll be discussing in this book, I believe. But then there are the *unknown unknowns*—future events, such as political changes,

technological innovations, or scientific breakthroughs that catch us entirely off guard.

In the next chapters, we turn to one of the great known unknowns: the puzzle of conscious experience. We will appreciate how this 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 may even be made of different substrates? A good place to begin is by simply appreciating the depth of the issue.

CHAPTER TWO

# 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. Although it may seem a stretch to claim that today's AIs are conscious, as they grow in sophistication, could it eventually feel like something to be them? Could synthetic intelligences 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.

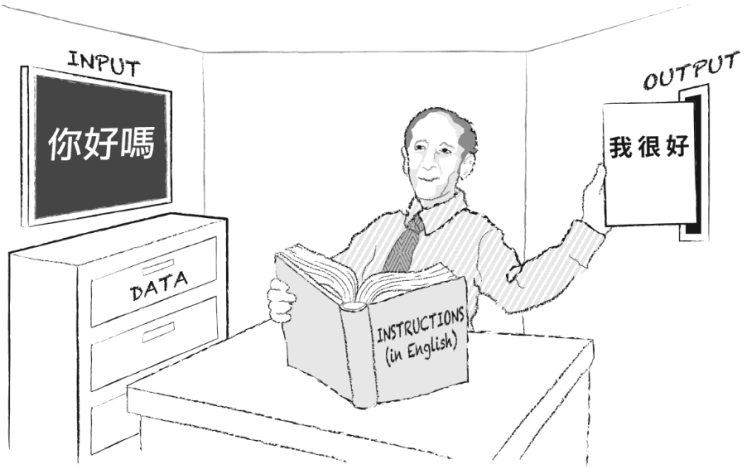
In the context of biological life, intelligence and consciousness seem to go hand-in-hand. Sophisticated biological intelligences tend to have complex and nuanced inner experiences. But would this correlation apply to nonbiological intelligence as well? Many suspect so. For instance, transhumanists, 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.[1] But as we shall see, this line of reasoning is premature. There may be no special androids that have the

## BIOLOGICAL NATURALISM

If biological naturalists are correct, then a romance or friendship between a human and an AI, like Samantha 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 any more 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 in the cloud, but that stream of data would not be you; it wouldn't have an inner life.

Biological naturalists suggest that consciousness depends on the particular chemistry of biological systems—some special property or feature that our bodies have and that machines lack. But no such property has ever been discovered, and even if it were, that wouldn't mean AI could never achieve consciousness. It might just be that a different type of property, or properties, gives rise to consciousness in machines. As I shall explain in Chapter Four, to tell whether AI is conscious, we must look beyond the chemical properties of particular substrates and seek clues in the AI's behavior.

Another line of argument is more subtle and harder to dismiss. It stems from a famous thought experiment, called "The Chinese Room," authored by the philosopher John Searle. Searle asks you to suppose that he is locked inside a room. Inside the room, there is an 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 rules (in English) that allows him to look up




Searle in the Chinese Room

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 consults his book, writes down Chinese symbols, and passes the card through a second hole in the wall.[5]

You may ask: What does this have to do with AI? Notice that from the vantage point of someone outside the room, Searle's responses are indistinguishable from those of a Chinese speaker. Yet he doesn't grasp the meaning of what he's written. 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.

Strictly speaking, this thought experiment argues against machine understanding, not machine consciousness. But Searle takes the further step of suggesting that if a computer is incapable of understanding, it is incapable of consciousness, although he doesn't always make this last step in his thinking explicit. For the sake of argument, let us assume that he is right: Understanding is closely related to consciousness. 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 also in an overall state of wakefulness and awareness.

So, is Searle correct that the Chinese room cannot be conscious? Many critics have zeroed in on a crucial step in the argument: that the person who is manipulating symbols in the room doesn't understand Chinese. For them, the salient issue is not whether anyone in the room understands Chinese, but whether the *system as a whole* understands Chinese: the person plus the cards, book, room, and so on. The view that the system as a whole truly understands, and is conscious, has become known as the "Systems Reply."[6]

The Systems Reply strikes me as being right in one sense, while wrong in another. It is correct that the real issue, in considering whether machines are conscious, is whether the 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. Consciousness is a feature

of highly complex systems, not a homunculus within a larger system akin to Searle standing in the room.[7]

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 percent of the brain's neurons, yet we know that it isn't required for consciousness, because there are people who were born without a cerebellum but are still conscious. I bet there's nothing that it's like to be a cerebellum.

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, because conscious systems are far more complex. The human brain, for instance, consists of 100 billion neurons and more than 100 trillion neural connections or synapses (a number which is, by the way, 1,000 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 Tinkertoy case. Even if consciousness is a systemic property, not all systems have it. This being said, the underlying logic of Searle's argument is flawed, for he hasn't shown that a sophisticated AI would lack consciousness.

In sum, the Chinese Room fails to provide support for biological naturalism. But although we don't yet have a compelling argument *for* biological naturalism, we don't have a knockout argument *against* it, either. As Chapter Three explains, it is simply too early to tell whether artificial consciousness is possible. But before I turn to this, let's consider the other side of the coin.