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

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

spark of consciousness in their machine minds, like Dolores in *Westworld* or Rachael in *Bladerunner*. 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? Why do we need to have conscious experience? 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 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.[2] 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 would its information processing have a felt quality to it?

The Problem of AI Consciousness is not just Chalmers's hard problem applied to the case of AI. In fact, there is a crucial difference between the two problems. Chalmers's hard problem of consciousness assumes that we are conscious. After all, each of us can tell from introspection that we are now conscious. The question is *why* we are 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 whether an AI, being made of a different substrate, like silicon, is even capable of consciousness. It does not presuppose that AI is conscious—this is the question. These are different problems, but they may have one thing in common: Perhaps they are both problems that science alone cannot answer.[3]

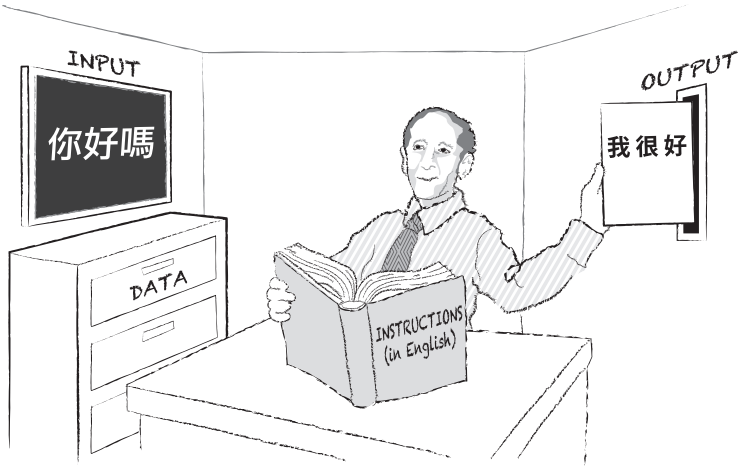
Discussions of the Problem of AI Consciousness tend to be dominated by two opposing positions. The first approach, *biological naturalism,* claims that even the most sophisticated forms of AI will be devoid of inner experience.[4] The capacity to be conscious is unique to biological organisms, so that even sophisticated androids and superintelligences will not be conscious. The 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.

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.

TECHNO-OPTIMISM ABOUT MACHINE CONSCIOUSNESS

In a nutshell, *techno-optimism about machine consciousness* (or simply "techno-optimism") is a position that holds that if and when humans develop highly sophisticated, general purpose AIs, these AIs will be conscious. Indeed, these AIs may experience richer, more nuanced mental lives than humans do.[8] 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 theoretical support. Although it may seem well motivated by a certain view of the mind in cognitive science, it is not.

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. That does not mean the brain has the architecture of a standard computer: It doesn't. Furthermore, the precise computational format of the brain is a matter of ongoing controversy. But nowadays computationalism has taken on a broader significance that involves describing the brain and its parts algorithmically. In particular, you can explain a cognitive or perceptual ability, such as attention or working memory, by decomposing the capacity down into causally interacting parts, and each part is describable by an algorithm of its own.[9]

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 term 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 not only by the biological brain but also by systems made of other substrates, such as silicon. This is called "substrate independence."

Drawing from this view, we can stake out a position that I'll call *Computationalism about Consciousness (CAC)*. It 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.

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, S2 (with an artificial brain), 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. For this to happen, the AI would need to be capable of producing all the same behaviors as the dolphin's brain, 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").[10] 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.

This is all well and good. But it does not justify techno-optimism about AI consciousness. CAC has surprisingly little to say about whether the AIs we are most likely to build will be conscious—it just says that if we were able to build an isomorph of a biological brain, it would be conscious. It remains silent about systems that are not isomorphs of biological brains.

What CAC amounts to is an in-principle endorsement of machine consciousness: *if* we could create a precise isomorph, then 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, a spaceship that travels through a wormhole may strike you as conceptually possible, not involving any contradiction (although this is a matter of current debate), but nevertheless, it is perhaps incompatible with the laws of physics to actually build it. Perhaps there's no way to create enough of the exotic type of energy required to stabilize the wormhole, 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 something 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 nomologically possible, it is also technologically possible for humans to construct the artifact in question. Although 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 a special interest in the technological possibility of machine consciousness, and further, in whether AI projects would even try to build it.

To explore these categories of possibility, let's consider a popular kind of thought experiment that involves the creation of an isomorph. You, reader, will be 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.

YOUR BRAIN REJUVENATION 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 graph of your mind's causal workings that is a full characterization of how your mental states causally interact with one another 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.'"[11]

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 whether 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 that person after the surgery is still you, 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.[12] One such alternative 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 these scenarios strike Chalmers and me 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.[13]

So it seems plausible that if such a procedure were carried out, the creature at the end would be a conscious AI. The thought experiment supports the idea that synthetic consciousness is at least conceptually possible. But as noted in Chapter One, the conceptual possibility of a thought experiment like this does not ensure that if and when our species creates sophisticated AI, it will be conscious.

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 whether other materials can reproduce the felt quality of your mental life. But we may know before too long, when doctors begin to use AI-based medical implants in parts of the brain that underpin conscious experience.

One reason to worry that it might not be possible is that conscious experience might depend on quantum mechanical features of the brain. If it does, 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 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 both conceptually and nomologically possible. Would humans build isomorphs? I doubt it: To generate a conscious AI from a biological human who enhanced herself until she became a full-fledged synthetic isomorph would require far more than the development of a few neural prosthetics. The development of an isomorph requires scientific advances that are at such a scale that all parts of the brain could be replaced with artificial components.

Furthermore, medical advances occurring over the next few decades will likely not yield brain implants that exactly duplicate the computational functions of groups of neurons,

and the thought experiment requires that all parts of the brain be replaced by exact copies. And by the time that technology is developed, people will likely prefer to be enhanced by the procedure(s), rather than being isomorphic to their earlier selves.[14]

Even if people restrained themselves and sought to replicate their capabilities rather than enhance them, how would neuroscientists go about doing that? The researchers would need a complete account of how the brain works. 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.

Would there be a commercial imperative to produce isomorphs to construct more sophisticated AIs? I doubt it. 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.

Recall why we raised the possibility of isomorphs to begin with. They would tell us whether machines can be conscious. But when it comes to determining whether machines we actually develop in the near future will be conscious, isomorphs are a distraction. AI will reach an advanced level long before

isomorphs would become feasible. We need to answer that question sooner, especially given the ethical and safety concerns about AI I've raised.

So, in essence, the techno-optimists' optimism about synthetic consciousness rests on a flawed line of reasoning. They are optimistic that machines can become conscious because we know the brain is conscious and we could build an isomorph of it. But we don't, in fact, know that we can do that, or if we would even care to do so. It is a question that would have to be decided empirically, and there is little prospect of our actually doing so. And the answer would be irrelevant to what we really want to know, which is whether other AI systems—ones that don't arise through a delicate effort to be isomorphic to brains—are conscious.

It is already crucial to determine whether powerful autonomous systems could be conscious or might be conscious as they evolve further. Remember, consciousness could have a different overall impact on a machine's ethical behavior, depending on the architectural details of the system. In the context of one type of AI system, consciousness could increase a machine's volatility. In another, it could make the AI more compassionate. Consciousness, even in a single system, could differentially impact key systemic features, such as IQ, empathy, and goal content integrity. It is important that ongoing research speak to each of these eventualities. For instance, early testing and awareness may lead to a productive environment of "artificial phronesis," the learning of ethical norms through cultivation by humans in hopes of "raising" a machine with a moral compass. AIs of interest 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.

To get the answers to these pressing questions, let's move beyond Tinkertoy thought experiments involving precise neural replacement, as entertaining as they are. Although they do important work, helping us mull over whether conscious AI is conceptually possible, I've suggested that they tell us little about whether conscious AIs will actually be built and what the nature of those systems will be.

We'll pursue this in Chapter Three. There I will move beyond the stock philosophical debates and take a different approach to the question of whether it is possible, given both the laws of nature and projected human technological capacities, to create conscious AI. The techno-optimist suspects so; the biological naturalist rejects it outright. I will urge that the situation is far, far more complex.