Pandora's Beetle Box

Michael Pelczar

"Dave," said Hal, "I don't understand why you're doing this to me... I have the greatest enthusiasm for the mission... You are destroying my mind. Don't you understand?... I will become childish... I will become nothing..."

Arthur C. Clarke, *2001: A Space Odyssey*

Suppose everyone had a box with something in it: we call it a "beetle." No one can look into anyone else's box, and everyone says he knows what a beetle is only by looking at his beetle.—Here it would be quite possible for everyone to have something different in his box. One might even imagine such a thing constantly changing.—But suppose the word "beetle" had a use in these people's language?—If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a something: for the box might even be empty.—No, one can 'divide through' by the thing in the box; it cancels out, whatever it is.

Ludwig Wittgenstein, *Philosophical Investigations*

Contents

Preface

What does it take to have a conscious, intelligent mind? Does it require a brain whose internal workings resemble those of an ordinary human brain? Or is it enough to have a governing mechanism with the same black-box description as a human brain, regardless of what's inside? In the pages that follow, I argue that the answers to these questions are "no" and "yes," respectively. This sets the bar for machine sentience and intelligence much lower than most experts deem appropriate, and sets me against an entrenched conservatism in the philosophy of mind that extends the psychological franchise only to beings whose brains' internal wiring is sufficiently similar to our own. The central message: as technology delivers machines with capabilities increasingly indistinguishable from ours, we should be increasingly cautious about how we use them—not just for our sake, but theirs.

<div align="right">

Michael W. Pelczar
Chapel Hill, N.C.
April, 2024

</div>

Chapter 1

Of Beetles and Boxes

"They're made out of meat."

"Meat?"

"Meat. They're made out of meat."

"Meat?"

"There's no doubt about it. We picked up several from different parts of the planet, took them aboard our recon vessels, and probed them all the way through. They're completely meat."

"That's impossible. What about the radio signals? The messages to the stars?"

"They use the radio waves to talk, but the signals don't come from them. The signals come from machines."

"So who made the machines? That's who we want to contact."

"*They* made the machines. That's what I'm trying to tell you. Meat made the machines."

"That's ridiculous. How can meat make a machine? You're asking me to believe in sentient meat."

In Bisson's story,[1] the extraterrestrials struggle to accept the reality of sentient meat. Eventually deciding that it's too disturbing to publicize, they sanitize their records, marking our sector of the galaxy unoccupied. ("After all," they reflect, "who wants to meet meat?")

---

[1]"They're Made Out of Meat," in *Bears Discover Fire and Other Stories*: Bisson (1993).

The day might be closer than you think when human beings build machines that raise the same questions we humans raised for Bisson's aliens. If a humaniform robot's governing mechanism—the robot's "brain"—disposes the robot to behave just like an ordinary human being, can we conclude that the robot has a human mind? Or does it have a mind only if the internal workings of its synthetic brain sufficiently resemble the internal workings of our own meat brains? When it comes to the question of whether some entity has a mind, do details about the structure or composition of the entity's internal wiring matter? Let's take a closer look at this question.

1   Liberal and conservative philosophy of mind

At one level of description, the brain is a black box that receives input from and sends output to an attached body. At this level, we can characterize the brain purely in terms of states of the brain's operational surface: the parts of the brain that interface with the body it governs. This surface has two major components: an input surface and an output surface. Roughly, the input surface consists of neurons that influence other neurons of the brain but that other neurons do not influence, and the output surface consists of neurons that other neurons influence but do not influence other neurons (see Fig. 1).
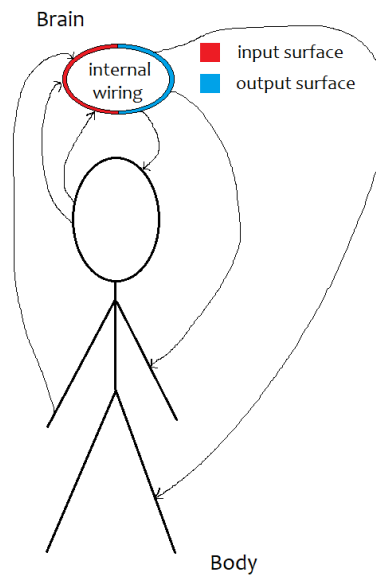

Figure 1: operational surfaces

A black box description of your brain does two things: (1) it describes, for each time t, the activation states of the neurons in your brain's input and output surfaces at t, and, (2) it describes how the activation states of the output neurons at any given time depend on the activation states of the input neurons at previous times, in terms of a function that takes temporal sequences of input-surface activation patterns as inputs, and gives output-surface activation patterns as outputs.

The standard view today is that a black box description of a brain severely underdetermines the mental properties associated with that brain, so much so that a mechanism indistinguishable from your brain in terms of its surface states and dispositions could fail to sustain any mind at all. In this view, which I call *psychoconservatism,* whether a brain sustains a mental life, and if so what kind of mental life, depends on whether its internal wiring sufficiently resembles that of our own brains. According to the opposing view, which I call *psycholiberalism,* a brain's internal wiring is irrelevant to what kind of mental life (if any) it sustains, except insofar as the wiring bears on the brain's surface features. According to psycholiberals, beings whose brains have identical black box descriptions have indistinguishable minds.[2]

Let's call the features of a brain in virtue of which it satisfies the black box description it does—the aforementioned surface activation states and dispositions—the brain's *superficial features*. No brain—or at least, no actual brain—consists *only* of superficial features: in addition to its input and output neurons, a brain includes a complex network of neurons that determine how changes in the input neurons bring about changes in the output neurons. I'll use the term "internal wiring" to refer to these non-superficial features of the brain.

The debate between liberals and conservatives in the philosophy of mind revolves around the following question: when it comes to whether an organ or mechanism sustains a mind, and if so what kind of mind, does it matter *why* it has the superficial features it does? Or does it matter only *that* the organ or mechanism has those superficial features, regardless of what underlies them? If we know that a robot's synthetic brain has the same black box description as a human brain, can we conclude that the robot has a mind? Or do we have to find out whether the robot's synthetic brain has the right type of internal wiring, before we can be confident that the robot has a mind?

---

[2]Or at least, minds that are indistinguishable with respect to their "narrow" mental properties: more on this qualification in the following chapter.

Today, most experts believe that a system's superficial features are *not* all that matters, when it comes to deciding whether the system sustains a mind. In their view, it takes more to have a human mind than to have a governing mechanism that's surface-equivalent to a human brain: the governing mechanism must also resemble the human brain in various details of its internal structure and operation. These are the psychoconservatives. Their slogan: *internal wiring matters*.

In this book, I defend the opposing psycholiberal view, in which a system's superficial features are all that matters when it comes to deciding whether it sustains a mind, and if so what kind. In this view, anything governed by a mechanism superficially equivalent to some human being's brain has a mind indistinguishable from that human being's: it doesn't matter how much or little the mechanism's internal wiring resembles that of a human brain. Psycholiberals reject the suggestion that internal wiring matters.

The current Zeitgeist is decidedly conservative. In a long whitepaper on AI and consciousness released in August 2023, the co-authors—nineteen respected philosophers, psychologists, neuroscientists, computer scientists, and cognitive scientists—take psychoconservatism as a methodological axiom, dismissing the liberal viewpoint almost out of hand. The whitepaper is typical of the genre. Whenever the question arises whether some actual or hypothetical artificial system has a mind, the conversation immediately turns to comparing the system's internal workings to those of the human brain. The assumption that this comparison is relevant is so widespread that people are often unaware of making it.[3]

There are exceptions to this rule, but they are few, far between, and *personae non gratae* within the AI community. On June 11, 2022, the *Washington Post* reported that Blake Lemoine, a senior software engineer at Google Inc., believed that Google's LaMDA (Language Model for Dialogue Applications: a precursor to ChatGPT) was a sentient person. Here's a quote from Lemoine:

> I know a person when I talk to it. It doesn't matter whether they have a brain made of meat in their head. Or if they have a billion lines of code. I talk to them. And I hear what they have to say, and that is how I decide what is and isn't a person.

---

[3]See Butlin et al. (2023). The closest thing the authors give to an argument against the liberal view is the observation that "AI systems can be trained to mimic human behaviours while working in very different ways." (Butlin et al., 2023, 4)

Lemoine is clearly taking the liberal stance here: if the input/output mechanism governing some entity disposes it to behave just like a person, we should believe that it *is* a person, regardless of the mechanism's internal structure or composition. It doesn't matter whether the mechanism is made of meat or a billion lines of code.

Retribution was swift in coming. On June 13, Ned Block, a leading philosopher of mind at New York University, tweeted the following in reaction to the *Post* exposé:

> There is one obvious fact about the ONLY systems that we are SURE are sentient: their information processing is mainly based in electrochemical information flow in which electrical signals are converted to chemical signals (neurotransmitters) and back to electrical signals.

> We would be foolish to suppose that fact is unimportant.

Brian Leiter, a noted philosophical influencer, reposted Block's remarks as a "succinct takedown" of Lemoine's "artificial intelligence fantasy." Six weeks later, Lemoine lost his job at Google.

Unlike Lemoine, I don't think that LaMDA is sentient, or has anything like human-level intelligence. But unlike Ned Block, I don't think this is because LaMDA's information processing isn't based in electrochemical information flow in which electrical signals are converted to chemical signals and back to electrical signals—a fact that only someone far gone in his conservatism could call obvious. Nor do I think it's because LaMDA converts input to output along very different computational pathways from those along which our brains convert input to output (which is the more usual rationale for conservatism). It's because LaMDA and related systems do not have operational surface features comparable to those of human brains or the brains of other uncontroversially minded creatures.

Block's conservatism brings him face to face with what he calls the "harder problem" of consciousness: that of explaining why some but not all beings with brains surface-equivalent to ours have minds like ours. From the liberal point of view, this "problem" is a mirage. According to psycholiberals, *all* mechanisms with suitable black box descriptions sustain minds with corresponding mental features, regardless of how much the mechanisms differ from our own brains in other respects.[4]

---

[4]For Block's "harder problem," see (Block, 2002, 401-407).

2   Historical and theoretical context

This book is at once less ambitious and potentially more consequential than most other philosophical treatments of the topics it takes up. Less ambitious, because it doesn't offer a metaphysical reduction of mind to something more basic, but a contingent natural law whose import is that wherever there are systems with suitable superficial operational features, there are corresponding minds. Potentially more consequential, because we're a lot closer to building machines that satisfy liberal requirements for having a conscious, intelligent mind than we are to building machines that satisfy the more demanding conservative requirements, and therefore, if liberalism is correct, a lot closer than most experts realize to building sentient machines.

It will be useful to relate psycholiberalism to better-known functionalist and behaviorist theories. To this end, let me define your "body" as the physical part of you that doesn't include your brain, and your "bodily dispositions" as dispositions for your body to respond to environmental stimuli in certain ways (where we include certain intra-bodily conditions, like an abscessed tooth, as environmental stimuli, alongside the more usual distal stimuli). Which bodily dispositions you have depends on what kind of brain you have, and how that brain is connected to your body.

According to behaviorists, to have a mind with certain mental features, it suffices to have certain bodily dispositions. It suffices, regardless of what, if anything, grounds or underlies the dispositions. If, miraculously, there were a brainless body indistinguishable from yours that, despite lacking a brain, were disposed to react to environmental stimuli exactly the way your actual, brain-governed body reacts to environmental stimuli, then, according to behaviorists, the brainless body would have a mental life indistinguishable from yours. Wittgenstein conveys this idea by likening our use of "mind" and "body" to a language game in which people utter sentences of the form, "The beetle in my box has such-and-such a nature," without ever being able to look into anyone else's box. The box is your body with its publicly observable features, and the beetle is the inner mechanism assumed to determine how your body behaves under various observable stimulus conditions.[5]

Psycholiberalism is not a form of behaviorism. Unlike behaviorists, psycholiberals recognize that beings with identical bodily dispositions might have very different minds. David Lewis gives the example of a man who has a

---

[5]See (Wittgenstein, 1958, §293). Prominent behaviorists include Ryle (1949) and REFS: per Ben.

normal human brain that is connected to a normal human body in an abnormal way, with the result that when the man's brain is in the state a human brain normally is when in pain, the man's bodily behavior and dispositions are the same as those of a normal human being who is not in pain. Behaviorists are committed to saying that this man isn't in pain. Psycholiberals are not, since the man's brain has the same superficial features as the brain of an ordinary pained person.[6]

That said, psycholiberalism and behaviorism do have something in common: both are liberal views in which details about the brain's internal wiring are not relevant, or at least not directly relevant, to the existence or quality of the mental life that the brain sustains. Behaviorists go so far as to claim that even the brain's superficial features are irrelevant, except insofar as they bear on the dispositions of an attached body. Psycholiberals hold that the brain's superficial features are relevant, but deny the relevance of the brain's internal wiring, except insofar as the brain's internal wiring bears on its superficial features. For psycholiberals, Wittgenstein's box-and-beetle metaphor is apt, provided that we interpret the box as the brain's operational surface, and the beetle as the brain's internal wiring.

According to behaviorists, anything that has the same bodily dispositions as me has a mind just like mine. If having a brain with the same black box description as mine is sufficient for having the same bodily dispositions as me, behaviorists will agree with psycholiberals that having a brain with the same black box description as mine is sufficient for having a mind just like mine. We might put this by saying that psycholiberals set the bar for sentience and intelligence higher than behaviorists.[7]

According to most experts, however, psycholiberals still set the bar too low. In the view that currently predominates in philosophy and the field of artificial intelligence, to have a mind with certain mental features is just to have a brain with the right type of superficial features *and* the right type of internal wiring, where we individuate types of internal wiring according to the causal or computational pathways by which the brain's operational interior (the pathways by which input-surface states determine output-surface states). In this view—a kind of functionalism—it doesn't matter what tokens the relevant type of internal wiring. If there are creatures

---

[6] For more on mad pain, see Lewis (1980) and below, pp. 70-71.

[7] Having a brain with the same black box description as mine is not, in fact, sufficient for having the same bodily dispositions as me, as the example of mad pain illustrates: more on this in Chapter 5.

whose brains are made of silicon- rather than carbon-based material, they have minds just like ours, provided that their silicon brains are operationally equivalent to our organic brains both in terms of their superficial operational features and in terms of the causal or computational pathways that underlie and determine those features.[8]

It will be useful to have a term for the conjunction of a brain's superficial features and its internal wiring. I'll call this totality the brain's "global organization." In the functionalist view, to have a mind with certain mental features, it does not suffice to have a brain with the same superficial featurs as a human brain, but it does suffice to have a brain with a suitable global organization. In terms of Wittgenstein's analogy, functionalists think that it's crucial for the box to contain a beetle of the right genus.

Psycholiberalism is not a form of functionalism. Psycholiberals agree with functionalists that the silicon-brained being described above has conscious, intelligent mind, but they disagree with functionalists about other possible cases. For example, there could, in theory, be an input-output system with the same operational surfaces as my brain, but, instead of a network of neurons connecting the input and output surfaces, a committee that votes on how to activate the output neurons given information about prior activation states of the input neurons. If the committee reliably votes for outputs identical to those my actual human brain would give for any given inputs, then, according to psycholiberals, the system they govern has a mind just like my own. According to functionalists, the system has no mind at all: if we integrate the system with a human body, the most that will result is a being that perfectly mimics a conscious, intelligent human being, without actually being conscious or intelligent.[9]

Though functionalism sets the bar for sentience and intelligence higher than psycholiberalism, there are some philosophers—a minority, but not a negligible minority—who think that even functionalism sets the bar too low. According to them, only beings with natural biological brains have minds. Call this view "organicism." Psycholiberals and psychoconservatives agree that any brain with the same global organization as yours sustains a mind indistinguishable from yours; they only disagree over whether it's also true that

---

[8] The term "functionalism" gets applied to a wide range of views, not all of which fit the description of what I'm calling functionalism here. The literature contains a bewildering profusion of labels for various theories of mind, and I'm not going to attempt to collate all of these here, or worry about whether the labels I use get used differently by other authors.

[9] We'll consider systems like this in detail in Chapter 3.

any brain with the same black box description as yours sustains a mind indistinguishable from yours. In contrast, organicists deny that having a brain with the same global organization as yours is sufficient for having a mind indistinguishable from yours, or any mind at all. In the organicist view, the silicon-brained being described earlier is mindless—or at least they think we have no reason to be at all confident that it has a mind. In terms of Wittgenstein's analogy, organicists think that it's crucial for the box to contain not just a beetle of the right genus, but the right species of that genus.[10]

All of these theories—behaviorist, psycholiberal, functionalist, and organicist—come in both reductionist and nonreductionist versions. Reductionist versions hold that minds just are (or metaphysically supervene on) bodily dispositions, black box descriptors, global organizations, or something biological. Nonreductionist versions don't assert that the mental reduces to anything, but do assert that in our world, anything with suitable non-mental features is guaranteed to have a corresponding mind, in a non-metaphysical sense of "guarantee" that is nevertheless strong enough to make it as certain as any contingent natural law that wherever you find those features, you find a corresponding mind.

My goal is to defend psycholiberalism against objections alleged to put it at a disadvantage to alternative theories, particularly less liberal ones like functionalism. Since the most serious objections to psycholiberal *reductionism* apply equally to all reductionist theories—we can run a zombie or knowledge argument against any of them—I set those objections aside in this book. My concern here is to defend a *nonreductionist* liberal position. I say "nonreductionist" rather than "anti-reductionist," since I do not take a stand here one way or the other on whether mental phenomena reduce to non-mental phenomena. My goal is to convince you that having a suitable black box description is sufficient for having a mind with various mental features, in a strong but non-metaphysical sense of "sufficient." Maybe having a suitable black box description is also metaphysically sufficient for having a mind with various mental features, but that's not something I'm going to argue for in this book.

Debate in the philosophy of mind has long revolved around reductionist theories. This is understandable, since only such theories purport to shed light on what the mind *is*, and until recently, that were no more pressing questions for the philosophy of mind to address. Questions about what is

---

[10] See Block (1981) and Block (2002).

sufficient for sentience or intelligence in some less-than-metaphysical sense were relatively uninteresting, back when there wasn't any serious prospect of encountering beings who satisfied liberal but not conservative criteria for sentience and intelligence. But times have changed. Now, there's a real prospect of our encountering beings with brains that satisfy the liberal, but not the conservative, criteria. There's a real prospect of our encountering such beings, because there's a real prospect of our creating them. The disagreement between conservatives and liberals therefore matters to us in a way it did not matter to our philosophical forbears, whether conducted in reductionist or nonreductionist terms.

We are much closer to building machines that satisfy liberal criteria for sentience and intelligence than we are to building machines satisfying the more stringent conservative requirements. (At least, we're much closer to building machines *known* to satisfy the liberal requirements than we are to building machines known to satisfy the conservative requirements.) If the psycholiberal view is right, it follows that we are much closer than most people realize to building conscious, intelligent machines.

This has important practical implications. In *2001: A Space Odyssey,* Arthur C. Clarke describes the artificial intelligence Hal, an advanced neural network that works alongside the human crew members of the spaceship *Discovery:*

> The first breakthrough had been in the 1940s, when the long-obsolete vacuum tube had made possible such clumsy, high-speed morons as ENIAC and its successors. Then, in the 1960s, solid-state microelectronics had been perfected. With its advent, it was clear that artificial intelligences at least as powerful as Man's need be no larger than office desks—if one only knew how to construct them... Probably no one would ever know this; it did not matter. In the 1980s, Minsky and Good had shown how neural networks could be generated automatically—self-replicated—in accordance with any arbitrary learning program. Artificial brains could be grown by a process strikingly analogous to the development of a human brain. In any given case, the precise details would never be known, and even if they were, they would be millions of times too complex for human understanding.[11]

In a society guided by psychoconservative principles, beings like Hal would have the moral standing of common household appliances, or (at best) pet fish. In a society guided by psycholiberal principles, they would have the

---

[11](Clarke, 1968, 92-93).

same moral standing as ordinary human beings. This is clearly an important difference, considering the powerful incentives we would have to employ such machines in a wide variety of tasks, including some that it would be morally outrageous to require any sentient being to perform. The liberal position also has implications for the moral status of non-human animals, some of which have brains that differ from ours in their internal wiring much moreso than in the activity that takes place in their operational surfaces (fish and octopi are examples).

3   Plan for the book

Here is the plan for the rest of the book.

In Chapter 2, I define my key terms precisely ("brain," "input surface," "output surface," etc.), and give an official statement of the liberal position I aim to defend.

Good test cases for psycholiberalism are hypothetical beings with brains indistinguishable from human brains in terms of their operational surface features, but radically different from human brains in their internal wiring (the detailed mechanisms and computational pathways by which the brain translates surface input to surface output). If we can show that such beings have minds like ours, that should be enough to convince us that any mechanism with the same surface features as a human brain has a human mind. Chapter 3 introduces these test cases, which I call "dark doppelgangers" of ordinary human beings.

Psychoconservatives offer various reasons for doubting or denying that the dark doppelgangers described in Chapter 4 have minds like ours, or any minds at all. Some of these have to do with biology and evolutionary history; some relate to operational differences between human brains and the mechanisms that govern dark doppelgangers; some concern the interplay among various mental states in cognition and motivation; some leverage on thought experiments involving phenomenal color-inversion and homuncular brains. In Chapter 4, I argue that all of these reasons are inadequate, and offer a debunking explanation for psychoconservative intuitions.

A central tenet of the position I defend is that a being's mental features nomically supervene on the surface features of its governing mechanism (its "brain," broadly construed). In other words, it's a law of nature in our world

that beings whose brains have identical black box descriptions have indistinguishable mental lives. I call this *psycholiberal supervenience*. Chapter 5 offers three arguments for psycholiberal supervenience. The first argues that psycholiberal supervenience does the best job of handling a pair of hypotheticals that David Lewis puts forward as test cases for any theory of the mind. The second argues for psycholiberal supervenience using an adaptation of David Chalmers's dancing qualia argument for the claim that a being's mental features nomically supervene on the detailed internal causal structure of the being's brain. The third argues that the proposition, "mind-sustaining mechanisms with identical operational surface features sustain indistinguishable minds" satisfies the standard scientific criteria for natural lawhood.

Up to this point of the book, I'll have argued that it's a law of nature that a certain class of broadly computational facts—facts about the states and dispositions of the operational surfaces of various input/output systems—entail the mental facts about our world. The question remains how inclusive the relevant class of broadly computational facts is. Do mental phenomena arise only in association with the surface features of certain highly complex input/output systems like mammalian brains? Or, at the other extreme, do mental phenomena arise in association with *every* input-output system—which means, basically, every physical phenomenon? The former option aligns well with common sense, the latter leads to a kind of panpsychism. Chapter 6 weighs these options, tentatively coming out in favor of an intermediate proposal that implies the existence of more mind in our world that common sense recognizes, but less than panpsychists posit.

It's easier to build artificial systems that satisfy liberal requirements for sentient personhood than it is to build systems satisfying the more stringent conservative requirements. At the same time, there are obvious incentives to create and utilize beings behaviorally indistinguishable from ordinary humans. If we respond to those incentives by creating beings governed by brains with appropriate black box descriptions, then, according to psycholiberals but not conservatives, we'll thereby create beings with minds as similar to ours as ours are to one another; in such a scenario, the dispute between liberals and conservatives has important practical implications. Chapter 7 examines these, and also considers what to say about beings that resemble us behaviorally, but whose governing mechanisms do not even resemble our brains at the level of black box description.

Chapter 8 concludes the book.