Martian Madness

Michael Pelczar

What does it take to have a conscious, intelligent mind? Does it require a brain whose internal workings resemble those of an evolved human brain? Or is it enough to have something with the same black-box description as a human brain, regardless of what's inside? 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. The central message: as we build machines with capabilities increasingly indistinguishable from ours, we should be increasingly cautious about how we use them—not just for our sake, but theirs.

1 Introduction

At one level of description, the brain is a black-box that receives input from and sends output to its environment. At this level, we can characterize the brain purely in terms of states of its operational surface. This surface has two major components: an input surface, consisting of the parts of the brain that the world outside the brain (the brain's environment) directly influences, and the output surface, consisting of the parts of the brain that directly influence the world outside. Typically, the locus of these influences is a body to which the brain is connected (see Fig. 1).[1]

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

[1] Strictly speaking, every part of a brain directly influences and gets directly influenced by the outside world; e.g., Pluto exerts a gravitational influence on each of your brain cells and vice versa. To accommodate this, we can define a brain's operational surfaces as the parts of it that directly affect or get directly affected by the brain's environment in ways that depend on those parts belonging to a brain.

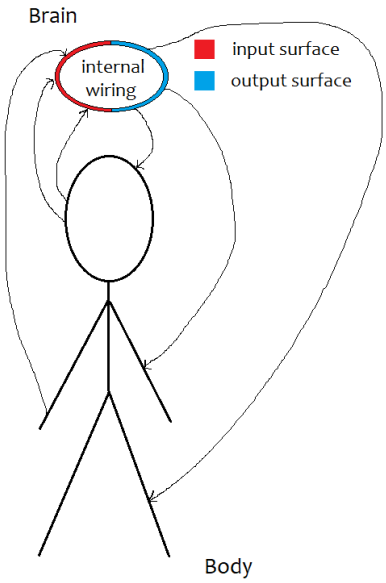

Figure 1: operational surfaces

Call the properties of a brain in virtue of which it satisfies the black-box description it does the brain's *superficial properties*. No brain—at least, no actual brain—has *only* superficial properties: in addition to the neurons constituting its input and output surfaces, a brain includes a complex underlying network of neurons that determines how changes in the input surface bring about changes in the output surface. I'll use the term "internal wiring" to refer to these non-superficial features of the brain.[2]

By an individual's "brain," I mean the smallest spatiotemporal part of the individual that is capable of sustaining the individual's mind with its intrinsic mental features: basically, the individual's smallest BIV-able component.[3]

---

[2]In neuroscientific terms, the neurons that constitute a brain's input surface are sensory neurons, and those that constitute a brain's output surface are motor neurons. The rest of a brain's neurons—those that constitute the brain's internal wiring—are so-called interneurons. In machine-learning terms, a brain's input surface is its input layer, its output surface is its output layer, and its internal wiring consists of its hidden layers.

[3]The qualification "intrinsic" is to set aside any "wide" mental features that an individual might have in virtue of how it relates to its social or physical environment, as in externalist accounts of mental representation like Burge (1977) and Millikan (1984).

We can agree about which part of you is your brain, without agreeing about which properties of it are such that having them is sufficient for possessing a mind with your mind's intrinsic mental features. (Analogously, we can agree about which part of a building is the smallest part capable of supporting its roof, even if we don't agree about which properties of this part are such that having them is sufficient for supporting the roof: maybe I think having the relevant part's geometry suffices, while you think having both the relevant part's geometry and density—but not just having its geometry—suffices.)[4]

Let's assume, plausibly, that your brain is the organ inside your skull, or some substantial portion thereof. Which properties of this organ are such that their instantiation is sufficient for the possession of a mind with your mind's intrinsic mental features?

My central claim is that your brain's superficial properties are sufficient for this. More generally, I claim that in our world, it's as certain as any natural law that systems with identical black-box descriptions sustain intrinsically indistinguishable minds. Call this *black-box supervenience:*

> It's nomically necessary that things with the same black-box description sustain intrinsically indistinguishable minds.[5]

Black-box supervenience is at odds with the prevailing wisdom, according to which a black-box description of a brain radically underdetermines the mental properties associated with that brain, so much so that something with the same black-box description as your brain could fail to sustain any mind at all. In this view, which I call *psychoconservatism,* whether an organ or mechanism sustains a mental life, and if so what kind of mental life, depends on whether it has the right kind of internal wiring.[6]

In the opposing view that I favor, 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 superficial properties. Psycholiberals uphold black-box supervenience.

---

[4]Here and throughout, by "sufficient" I mean *nomically* sufficient; more precisely, by "p is sufficient for q," I mean that it's as certain as any natural law that if p, then q.

[5]If two things both fail to sustain any mind, I count this as a degenerate case of sustaining indistinguishable minds.

[6]Butlin et al. (2023) is a recent statement of conservative orthodoxy.

It'll be useful to have a term for the conjunction of a brain's superficial features and its internal wiring. Let's call this the brain's "network-wide organization." In the standard conservative view, to have a mind like yours, it does not suffice to have a brain with the same superficial features as your brain, but it does suffice to have a brain with the same network-wide organization.[7]

Psycholiberalism says that any system with the same superficial properties as a given mind-sustaining system *S* sustains a mind intrinsically indistinguishable from the one that *S* sustains, and a natural extension of this is that systems with suitably similar superficial properties sustain correspondingly similar minds. However, psycholiberalism doesn't say anything about which systems sustain minds in the first place; e.g., it doesn't say whether insects or microbes have minds. This is a hard question that I don't try to answer here.

The psycholiberalism I defend is a nonreductionist theory.[8] This puts it at the margins of mainstream philosophy of mind, which has long revolved around reductionist theories—understandably, since only such theories purport to tell us what the mind *is*, and until recently, there were no more pressing questions for philosophers of mind to address.

But times have changed. Recent technological advances mean that there's now a real prospect of our creating beings that satisfy liberal, but not conservative, criteria for having a conscious mind. This gives the debate between conservatives and liberals a practical importance it never had before, independent of whether mental properties reduce to (rather than arise from or nomically supervene on) relevant operational or organizational properties.

The paper proceeds as follows. §2 reviews David Lewis's well-known cases of "mad pain" and Martian pain. This sets up for §3, which argues that psycholiberalism is uniquely qualified to give the right verdicts on these cases. The argument assumes (with Lewis) that there could be creatures who have pain, despite having brains radically unlike ours in their internal wiring. §§4-8 address various psychoconservative challenges to this assumption. §9 offers some speculation on the origins of psychoconservative intuitions. §10 concludes.

---

[7] Chalmers calls the functionalist sufficiency claim the "principle of organizational invariance": see (Chalmers, 1996, 248-49).

[8] I say "nonreductionist" rather than "anti-reductionist," since I take no stand here on whether mental phenomena reduce to non-mental phenomena.

2 Mad pain and Martian pain

David Lewis describes a "madman" whose brain is exactly like yours, but whose bodily behavior and dispositions are very unlike yours, due to how the madman's brain is connected to his body. For example, at a time when you are in pain, the madman's brain is in exactly the state your brain is in, but because of how the madman's brain is connected to his body, his bodily behavior and dispositions are indistinguishable from someone who is pain-free (see Fig. 2).[9]

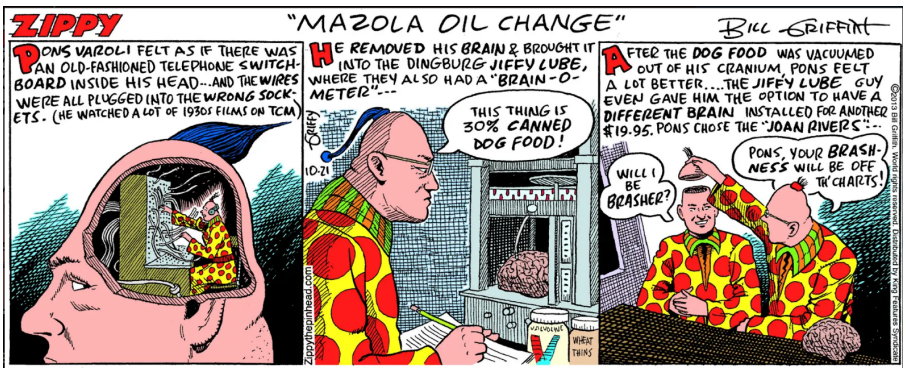

Figure 2: Lewisian Madness

Lewis also describes a Martian:

> [T]here might be a Martian who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its physical realization. His hydraulic mind contains nothing like neurons. Rather, there are varying amounts of fluid in many inflatable cavities, and the inflation of any one of these cavities opens some valves and closes others. His mental plumbing pervades most of his body—in fact, all but the heat exchanger inside his head. When you pinch his skin, you cause no firing of C-fibers—he has none—but rather, you cause the inflation of many rather smallish cavities in his feet. When these cavities are inflated, he is in pain. And the effects of his pain are fitting: his thought and activity are disrupted, he groans and writhes, he is strongly motivated to stop you from pinching him and to see to it that you never do it again. In short, he feels pain but lacks the bodily states that either are pain or accompany it in us.[10]

[9](Lewis, 1980, 216-219).

[10](Lewis, 1980, 216).

Lewis's description leaves some doubt as to the Martian's detailed anatomy. Let's assume that the Martian's body (the part of him that his hydraulics govern) is indistinguishable from an ordinary human body (two arms, two legs, a pair of eyes, etc.). Let's also assume that his hydraulics communicate with his body's sensory surfaces and motor anatomy via nerve fibers in the same way our brains communicate with our bodies' sensory surfaces and motor anatomy. The Martian is just like us, except for what he has in lieu of a human brain.

Lewis clearly intends for the Martian's hydraulics to differ greatly from a human brain, but in what respects? Do the hydraulics replicate the neuron-level functionality of a human brain, so that the hydraulic system performs the same computations as a human brain when converting sensory input to motor output, only using different physical processes to do so (e.g., fluid pressure gradients instead of ion gradients)? If so, then the Martian's hydraulic system is much the same as what you'd get by replacing each natural neuron of a human brain with a silicon chip or vacuum-tube assembly that replicates the functionality of the replaced neuron.[11]

However, I want to stipulate, in accordance with what I believe are Lewis's intentions, that the Martian's governing mechanism differs from a human brain more radically than this. It's not that the hydraulic system instantiates the same fine-grained causal structure as a human brain in a different medium, or using different physical processes; rather, the system converts surface input to surface output along causal and computational pathways utterly unlike those by which human brains convert input to output.

The clearest way to carry out this stipulation is by supposing that the Martian's hydraulic system works like a gigantic database or lookup table, like an automated Excel workbook with many spreadsheets. Each spreadsheet's first column has a large number of entries, in the form of alphanumeric strings. Each string is uniquely associated with a possible state of some Martian body's sensory surfaces (rods and cones, stereocilia, taste buds, Merkel cells, etc.). At each time increment, the computer running the workbook receives a pattern of signals from the body's sensory surfaces (via receivers feeding into the computer). This pattern gets translated into the string associated with that

[11] As in David Chalmers's "fading qualia" thought-experiment: (Chalmers, 1996, 253-63).

pattern in the computer's database. The computer finds the string in the first column of the currently open spreadsheet. In the adjacent second-column cell, there is another alphanumeric string, *S*. This string, like all the others in the second column, is associated with a possible state of the connected body's motor anatomy. The computer now sends out to the body (via transmitters connected to the computer) a signal associated with *S*, causing the body's motor anatomy to enter the relevant state (and thus move in certain ways, or, as the case may be, remain motionless). The computer now opens a new spreadsheet, whose number is indicated in the third column (adjacent to the cell containing *S*), and closes the current sheet. And so forth.

The twist is that the spreadsheets have been written in such a way that by integrating the workbook with a Martian body as described, the body engages in the same behavior, and has the same dispositions, as it would if it were governed by a human brain.

An agent governed by a system like this is what's known in the literature as a "Blockhead." The system (receivers, transmitters, and workbook) has the same black-box description as a human brain. The receivers that receive signals from the body constitute the system's operational input surface, and the transmitters that send signals to the body constitute the system's operational output surface. The activation states of these surfaces over time are the same as the activation states of some ordinary human brain, and the dispositions of the system's output surface to be in certain states given prior states of the system's input surface are also the same as in some ordinary human brain.[12]

The difference between Blockheads and humans is in their governing mechanisms' internal wiring. The automated workbook that constitutes the internal wiring of a Blockhead's governing mechanism is radically unlike the complex of neurons that constitutes the internal wiring of a human brain. We know this, because in order for an automated workbook to power a system with the same black-box description as a human brain, each of its spreadsheets would have to contain more rows than there are particles in the known universe.[13]

---

[12] Blockhead cases originate with (Block, 1981, 19-21); see also (Searle, 1984, 28-41), (Braddon-Mitchell and Jackson, 2007, 114-19), and (Kirk, 1974, 53-54).

[13] There are millions of points at which nerve fibers penetrate a human skull: millions of channels for signals to pass between the body and brain, and the same number of computer- or brain-side receivers. Even if each channel has only two available states (on/off), there are a

Stipulating that the Martian—Marvin, to give him a name—has a Blockhead mechanism in lieu of a human brain is one way to ensure that the internal wiring of his brain is radically unlike that of a human brain. For a more realistic thought-experiment, we could stipulate that Marvin has, in lieu of a human brain, a neural network akin to that which drives ChatGPT or the artificial intelligence HAL in Arthur C. Clarke's *2001: A Space Odyssey*.

It's actually debatable how unlike our brains such networks are—this is an area of ongoing research—but they are certainly not *known* to bear much structural or operational resemblance to naturally evolved brains (hence all the recent debate about "AI transparency"). Supposing we were to construct a neural net whose input layer received signals from the sensory surfaces of a human body (or an artificial body with the same environmental sensitivities as a human body), and whose output layer sent signals to the motor anatomy of the same body, we might train the network until it was capable of autonomously governing the body exactly as a natural human brain would govern it. Yet it might be practically impossible for us to know whether the network converts afferent input to efferent output along the same computational pathways as a natural human brain, or by totally different pathways. Agents with artificial neural networks in lieu of human brains would therefore raise many of the same questions as their more far-fetched Blockhead brethren. That said, let's assume that Marvin's governing mechanism is a Blockhead mechanism, since this will give my opponents their best chance of resisting the arguments that follow.

3 Psycholiberalism: the squeezing argument

The madman has pain, but we can't account for this along behaviorist lines by saying that he engages in behavior and has bodily dispositions characteristic of someone in pain, since he does not engage in such behavior or have such dispositions. So what does account for the madman's pain? Lewis reasonably accounts for it by reference to the fact that the madman's brain is in the same state *our* brains are in when we're in pain.

---

total of 2^{millions} possible signal patterns. Since each sheet in a Blockhead workbook has one row for each possible signal-pattern, each must have 2^{millions} rows in order for the system to work. But there are only about 10^{86} particles in the universe, and $10^{86} <<< 2^{\text{millions}}$.

According to Lewis, Marvin also has pain. Though Marvin engages in behavior and has bodily dispositions characteristic of someone in pain, we can't account for his pain by reference to this fact, since, as the case of the madman shows, a being's behavior and dispositions are unreliable guides to its mental states. (We can imagine that when the madman's behavior and bodily dispositions are identical to those of someone in pain, he has no pain.) So what accounts for Marvin's pain?[14]

Lewis doesn't try to account for Marvin's pain by saying that Marvin's brain is in the same state our brains are in when we're in pain. Perhaps he thinks Martian hydraulics are too unlike our grey matter to support such an account. Instead, Lewis says that Marvin is in pain because his hydraulics are in the state that Martian hydraulics normally are in when a Martian engages in the kind of behavior and has the kind of bodily dispositions that are typical of someone in pain.[15]

Marvin and the madman have this in common: both have bodies governed by mechanisms that are in the states that normally occur in the governing mechanisms of individuals belonging to their respective species when those individuals' bodies engage in the kind of behavior, and have the kind of dispositions, that are typical of someone in pain. It is this, according to Lewis, that explains why both Marvin and the madman are in pain.

This is also how Lewis accounts for the pain of a mad Martian whose hydraulics are connected to his body in a non-standard way, so that when his hydraulics are in the state that a Martian's hydraulics are normally in when a Martian is in pain, his bodily behavior and dispositions are those typical of a Martian who's not in pain. The mad Martian is in pain, says Lewis, because his hydraulics are in a state that *normally* underlies pain-behavior and pain-dispositions in members of the population to which the Martian belongs (the Martian species).[16]

---

[14] Lewis actually holds that Martian pain is pain in a somewhat different sense of "pain" from human pain. This is a questionable diagnosis, since whatever reasons we have to think that Marvin has pain appear to be reasons to think he has pain in the same *ouch-that-hurts* sense we do. I won't press this issue, since, as we'll soon see, Lewis's account suffers from far more serious implausibilities.

[15] (Lewis, 1980, 219).

[16] (Lewis, 1980, 220).

But what about a mad Martian who belongs to no population? (Perhaps he just popped into existence spontaneously, or maybe he's the very first member of his species and a practicing anti-natalist.) According to Lewis, "[o]ur only recourse is to deny that [this] case is possible."[17]

Denying the possibility of solitary mad Martian pain is a desperate move, especially coming from someone, like Lewis, who acknowledges the possibility of pain in a mad Martian who *does* belong to a population of Martians. If a mad Martian who belongs to a population of Martians can have pain, why can't a mad Martian who doesn't? Lewis's account turns pain into a relational property that one can have only if one has the right kind of connection to other beings similar to oneself. It is implausible that pain is such a property.

Relativizing mental states to populations commits Lewis to further implausibilities. He considers a subpopulation of humans in whom the brain states that sustain pain-related behavior and dispositions in other human beings instead sustain thirst-related behavior and dispositions. Lewis says that there is no determinate fact of the matter about whether the members of this subpopulation have sensations of pain or thirst when their brains are in the relevant state. This is a strange thing to say. On the face of it, the difference between feeling thirsty and having back pain is about as determinate as it gets.[18]

Is there a way we can say unequivocally that *both* the madman *and* the Martian have pain? Can both mad pain and Martian pain be pain, in the ordinary, determinate sense of the word? Can a being who belongs to no population, like a solitary mad Martian, have pain?

Yes. The Martian, the madman, and I all have this in common: our brains have identical (or relevantly similar) black-box descriptions, i.e. operational surfaces with identical (or relevantly similar) states and dispositions. Well, in the psycholiberal view, the brain's operational surface properties—its nature *qua* black box—are what determine the mental features of the mind that the brain sustains. These properties are identical (or relevantly similar) across the normal human being in pain, the victim of Mad Pain, and the suffering Martian. So psycholiberalism says that it's the same pain in all cases.

---

[17] (Lewis, 1980, 221).

[18] For Lewis's discussion of this case, see (Lewis, 1980, 220).

The madman acts just like someone who's not in pain, but he is in pain. His odd behavior is due to the abnormal effects his brain's outputs have on his body. The Martian's Blockhead mechanism doesn't convert input to output along pathways anything like those along which a human brain converts input to output, but he is in pain. His behavior arises from the same pattern of operational surface activity as normal human behavior, even though his brain's internal wiring is very different from that of a human brain.

Psycholiberalism also correctly diagnoses mad Martian pain, where a Martian's brain (Blockhead mechanism, neural net, or whatever) gets re-connected to the Martian's body in such a way that the Martian's behavior and dispositions are indistinguishable from someone who is having pleasure and no pain, even though the Martian's brain is in the state that a Martian's brain normally is in when having pain and no pleasure.

Last but not least, psycholiberalism accounts for the case that Lewis by his own admission cannot: a mad Martian who has pain despite belonging to no population. Since the superficial features of a mad Martian's brain do not depend on whether the Martian belongs to a population of similarly-constituted beings, solitary mad Martian pain is perfectly possible, in the psycholiberal view.

What do we find in cases of mad Martian pain that's distinctively also present in cases of normal human pain, mad human pain, and non-mad Martian pain? A governing mechanism with certain operational surface features. The black-box features of the normal man's brain = the black-box features of the madman's brain = the black-box features of the non-mad Martian's brain = the black-box features of the mad Martians' brains. Identical black-box features are the common thread throughout.

What if we stipulate that the Martian is *not* governed by a mechanism with a black-box description similar to that of a normal human brain? This new Martian—call her "Maya"—has, in lieu of a human brain, a governing mechanism that differs radically from our brains not only in terms of inner wiring but also in terms of superficial properties, so that its black-box description is unlike that of our brains or any brain known to us to be capable of causing or manifesting pain.

Unlike Marvin, we have no reason to think that Maya has experience anything like ours, or, depending on how radically Maya's governing mechanism differs from our brains, any mind at all. If, for example, the input surface of Maya's governing mechanism contains dramatically fewer nodes than the input surface of a human brain, perhaps receiving signals from sensory surfaces that are proportionally coarse-grained relative to human sensory surfaces, there's no reason to think Maya has perceptual experiences similar to ours, even if the relative coarseness of her physiology somehow (but how?) fails to prevent her from having bodily dispositions just like ours.

Suppose we discover humanoid beings whose skulls are full of gravel instead of brains. The gravel doesn't have the same kind of input-output architecture as our brains or any brain known to sustain a mind, let alone the same kind of internal wiring, nor is it part of a larger mechanism that has that kind of input-output architecture; yet, in spite of this, the gravel somehow manages to dispose the humanoids to behave just like they would if they had human brains instead of gravel in their heads.[19]

There's no more reason to think that these humanoids have minds than there is to think that piles of gravel do. But Maya is essentially just a gravel head. So we have no reason to think she has a mind either.

Most philosophers would agree. But most philosophers would say that we have equally little reason to think that Blockheads like Marvin have minds. Let's address this now.

4 The conservative challenge

I've argued that psycholiberalism is uniquely positioned to achieve a state of reflective equilibrium in our thinking about mad pain, Martian pain, and permutations thereof. This is a strong argument in favor of psycholiberalism, and against psychoconservative orthodoxy, provided that both the madman and Marvin have pain. Lewis clearly thinks that both do, and as far as the madman is concerned, he is surely right. (You can't relieve a person's pain just

[19] It's unclear that this is nomically possible, but let's suppose for the sake of argument that it is; maybe God has decided that wherever He finds a head full of gravel, he'll endow the body attached to it with normal human bodily dispositions, which He knows how to do without violating any natural laws.

by reconnecting his brain to his body in a way that results in an absence of pain-behavior or pain-related bodily dispositions.) But what about Marvin? Why should we think that he has pain?

Lewis doesn't give his reasons for attributing pain to Martians, but if pressed for a reason, I suspect he'd say that everyday norms of pain attribution commit us to thinking that Marvin has pain.[20]

It's a norm of everyday pain attribution that it's appropriate to attribute pain to anything that behaves in ways we normally take as evidence that someone is in pain, unless we have some good reason to withhold the attribution. Hilary Putnam calls this a "methodological directive":

> If some organism is in the same state as a human being in pain in all respects *known* to be relevant, and there is no reason to suppose that there exist *un*known relevant respects, then don't postulate any.[21]

More generally, the following seems like a reasonable rule:

> If everything about X's behavior suggests that X has a mind with mental feature ϕ, then we should believe that X has a mind with ϕ, unless we have some good reason to doubt that X has a mind with ϕ.

Even conservative philosophers of mind will agree that if something's behavior suggests that it's in pain, it's unreasonable to doubt that it's in pain absent *some* legitimate defeater of the proposition that the thing is in pain. Otherwise, we'd be free to doubt other people's pain whenever it was convenient to do so. The only question is: what counts as a defeater? More narrowly: what counts as a defeater of an attribution of pain to a being that, like Marvin, acts and is disposed to act just like it's in pain, and does some such defeater defeat the claim that Marvin has pain?

According to psychoconservatives, details about the internal wiring of Marvin's governing mechanism defeat the claim that Marvin has pain or any other mental state. Which details? There is curiously little agreement among conservatives about this; we consider some of their suggestions below. But first, let me address a prior issue.

---

[20] Regarding Martian pain, Lewis just says that a "credible theory of mind had better not deny the possibility of Martian pain." (Lewis, 1980, 217)

[21] (Putnam, 1975, 340).

Many psychoconservatives seem to think it's so obvious that a Blockhead like Marvin is mindless that the onus is on those who suggest otherwise to justify their stance. This raises a question of the burden of proof. Is it on liberals, to give reasons for thinking that Marvin has a mind like ours? Or is it on conservatives, to give reasons to doubt that Marvin has a mind?

It is the psychoconservatives who bear the burden. The key to seeing this is that *we don't know a priori that Marvin lacks a mind*. We don't know this *a priori,* because we don't know *a priori* that our own brains don't work like Marvin's hydraulics. Conservatives who say that Marvin is mindless must therefore take themselves to have some *a posteriori* basis for that claim.[22]

To be clear: that there is no valid *a priori* inference from, (1) "So-and-so has a mind" to (2) "So-and-so has a brain with such-and-such internal wiring" is no proof that it's possible for something to have a mind without having a brain with such-and-such internal wiring. There's no valid *a priori* inference from, "So-and-so has a glass of water" to "So-and-so has a glass of H_2O," but that doesn't mean it's possible for someone to have a glass of water without having a glass of H_2O. The point is that the unavailability of a valid *a priori* inference from (1) to (2) puts the burden on conservatives to provide some evidence for their claim that having a brain with the right kind of internal wiring is necessary for having a human mind.

Let's consider various conservative attempts to do just that: various ways they've challenged the idea that a Blockhead like Marvin has a mind.

5 Digital vs. analog

Some authors make much of the fact that existing artificial intelligences operate on a discrete or "digital" rather than continuous or "analog" basis, arguing that human brains, unlike (e.g.) Blockhead mechanisms, are analog systems, and that this is crucial to their mind-sustaining powers.[23]

---

[22] As noted above, we do have *a posteriori* evidence that *our* brains aren't Blockhead mechanisms, namely that this would require them to contain more matter than exists in the known universe. But the relevant question here is not whether our brains actually work like a Blockhead mechanism, but whether we would have the minds we do if they did. *A priori* reflection can't yield a negative answer to this question either: if it could, it could also answer the question whether our brains do in fact work like Blockhead mechanisms.

[23] See (Penrose, 1989, 405-450) and (Dyson, 2015, 85-98).

The difference between digital and analog systems is that digital systems achieve what they do using processes or algorithms described by functions defined over a countable number of possible inputs, whereas analog systems use processes or algorithms described by functions defined over a continuum of possible inputs.

The first thing to say about this is that we don't know that our own brains are analog systems. We don't know that *any* system existing in our universe is analog, since it's an open question whether actual spacetime is discrete or continuous. And even if spacetime is continuous, it's a further open question whether the physical phenomena it contains have a discrete or continuous structure. It would be strange if we could settle these questions just by referring to the fact that we have minds.[24]

Supposing, for the sake of argument, that spacetime is continuous, and that a human brain has an analog structure in virtue of which it's capable of undergoing arbitrarily small changes, it's doubtful that more than a finite subset of those changes are relevant to our mental lives. Since our cognitive powers are not infinitely sharp, and our perceptual powers not infinitely discriminating, it's hard to see why our brains would have to implement continuous functions in order to perform the cognitive and perceptual tasks they do.[25]

One likely source of confusion here is that people often use "analog" to mean, "most usefully represented by a continuous mathematical function." But a phenomenon that is most usefully represented by a continuous function might be a discrete phenomenon. For example, if you want to describe the growth rate of a population of bacteria in some medium, it might be most useful to do so with an equation that represents the size of the population as a continuous function of time. But the process being described is a discrete one, with the population only ever increasing by an integer number of bacteria. Likewise, even if the processing that occurs in our brains is most usefully represented by continuous functions, it doesn't follow that the processing itself is continuous.

---

[24] For arguments that spacetime is discrete, see (Smolin, 2001, 95-124) and Mäkelä (2011).

[25] In machine-learning terms, the point is that the only relevant differences between a perceptron and a sigmoid neuron are also differences between a perceptron and a neuron described by a suitably fine-grained step-function approximation of a sigmoid function.

At a more basic level, the suggestion that only analog systems can support minds is open to the criticism that there's no evident reason to think it's true. On the face of it, we have no more reason to think that our minds depend on spacetime being continuous than on its having a gently curved geometry. It's not hard to imagine beings in a flat or toroidal spacetime having brains with the same superficial or network-wide properties as our brains. If the spatiotemporal curvature of our brains contributes something critical to their mind-sustaining powers, it's up to those who say so to explain how. Likewise for those who say that the (alleged) continuity of our brains contributes something critical to their mind-sustaining powers. To my knowledge, nobody has explained how this might be.

6 Operational efficiency

Some psychoconservatives suggest that the mechanism governing Marvin is inefficient in a way that prevents it from sustaining a mind. Specifically, they suggest that the Marvin mechanism relies on an inefficient search strategy in a way that's incompatible with true intelligence.[26]

For example, according to Alan Newell and Herbert Simon, "the task of intelligence . . . is to avert the ever-present threat of the exponential explosion of search."[27] By "exponential explosion of search," Newell and Simon are alluding to input-output algorithms with the following property: in order for the algorithm to give its output for an input of size i (for some integer $i > 1$), the algorithm has to be applied n^i times (for some $n > 1$). For example, suppose A is an algorithm that takes strings of `0`s and `1`s as inputs, and gives other such strings as outputs. A is an "explosive" algorithm if, for all i, it takes 2^i applications of A for A to deliver an output for an input of a string i digits long.

Marvin's mechanism does rely on an explosive search strategy. This is due to the fact that the number of spreadsheets that the mechanism requires to

[26] (Block, 1981, 38) attributes this suggestion to Daniel Dennett.

[27] (Newell and Simon, 1979, 123), quoted in (Block, 1981, 38). Newell and Simon also make the more general claim that intelligence is essentially the avoidance of prohibitively costly problem-solving strategies (Newell and Simon, 1979, 121); this is implausible, since it implies that intelligence couldn't exist in an environment free from resource constraints.

function reliably up to a given time $t + 1$ is exponentially greater than the number of spreadsheets it requires to function reliably up to t. At each time increment i, the algorithm has to choose from among n^i spreadsheets to open next, since there are n^i possible i-string sequences of sensory inputs requiring an appropriate motor response (where n is the number of possible synchronic states of the mechanism's input surface).[28]

Explosive algorithms are highly impractical, given normal constraints on time and computing power. For example, even if it takes a computer just one picosecond to perform each iteration of algorithm A, it would take the computer about 30 billion years to give an output for a 100-digit input.

The class of functions computable in polynomial time is the class of functions that describe algorithms that are *not* impractical or unfeasible in this way. Computability in polynomial time is the efficiency analog of Turing computability: just as Turing computability formally captures the intuitive idea of a function whose outputs can be determined mechanically (i.e., without any insight or creativity), computability in polynomial time formally captures the intuitive idea of a function whose outputs can be determined feasibly or efficiently (i.e., without consuming an unrealistic amount of resources, or taking an unrealistic amount of time). An algorithm is executable in polynomial time just in case for an input of size n, it takes the algorithm no more than n^k steps to give an output, where k is some positive constant. For example, suppose B is an algorithm with the same input and output as our earlier algorithm A; however, unlike A, it takes B only i^2 iterations to give an output for an input of length i. Then for a 100-digit input, B gives an output in one tenth of a

[28] The mechanism might get lucky sometimes, finding the sheet it's searching for near the top of the pile, but, barring a highly improbable statistical fluke, on average it will have to dig exponentially deeper at each stage before it finds the sheet it's searching for. (Another way to appreciate the explosive nature of the Blockhead algorithm is to imagine it operating on a single spreadsheet whose first column has entries for all possible *sequences* of synchronic sensory input states. Assuming n possible states, the first n entries of the column will be single alphanumeric strings, the next n^2 entries will be ordered pairs of strings, the next n^3 entries will be ordered triples of strings, etc. To find a match for the initial incoming tranche of sensory input, the algorithm has to search through the first n entries; to find a match for the initial two-tranche sequence of inputs, it has to search through the next n^2 entries; to find a match for the initial three-tranche sequence of inputs, it has to search through the next n^3 entries, etc.)

nanosecond when we run it on the picosecond computer.[29]

The virtue of computability in polynomial time is entirely one of efficiency. If computer scientists weren't limited by the speed and energy requirements of the machines they programmed, they'd have no reason to devise polynomial-time algorithms to achieve things that they could already achieve with explosive algorithms. That would be like devising energy-efficient appliances in a future where cold fusion provides a limitless supply of clean energy.

Thus the only argument in favor of making polynomial computability a necessary condition for intelligence boils down to an appeal to efficiency.

But how does the relatively low efficiency of the mechanism that governs Marvin's body suggest that Marvin lacks a mind? Why doesn't it just show that Marvin has a mind sustained by a relatively inefficient mechanism—inefficient, that is, relative to ordinary human brains? (Or, more cautiously: that if Marvin has a mind, it's grounded in a relatively inefficient mechanism.)

We can imagine a world whose natural laws differ from our world's in such a way that it takes much more energy in that world to power a system with the structural and operational features of a human brain than one with the structural and operational features of Marvin's hydraulics. Do conservatives want to say that in such a world, human beings lack minds? Presumably not. Rather, they'll say that in the envisioned world, it takes much more energy than in our world to sustain a human mind.

More generally, it's hard to see why a system's efficiency in achieving the tasks it does should bear on whether the system sustains a mind. This is particularly so in view of the fact that there's no level of efficiency that stands out as being the minimum that we should require for genuine intelligence.

After all, even though our actual neural architecture is more efficient than Marvin's workbook architecture, it's presumably not the most efficient possible architecture. Evolution by natural selection finds efficient solutions, but not, in general, the most efficient in principle. We can imagine beings who have the same problem-solving capacities we do, in virtue of having brains with a much more efficient neural architecture than ours. As computer science and robotics advance, we might construct such beings ourselves. Our own species

[29] The idea of equating the intuitive notion of computational efficiency or feasibility with computability in polynomial time originates with Alan Cobham: see Cobham (1965).

might evolve into descendants with brains that work more efficiently than our current brains, or it could be that intelligent life has arisen elsewhere in the universe, under conditions that led to the evolution of beings with brains more efficient than ours.

Such beings need not be superior to us in terms of their problem-solving abilities. It could be that the *only* respect in which they surpass us is in the operational efficiency of their brains, which might be indistinguishable from our brains at the level of black-box description.

Presumably it would be a mistake to say that such beings lack intelligence, simply because their brains are more efficient than ours. It would also be a mistake, of course, to say that *we* lack intelligence because our brains are less efficient than theirs. But if our brains' inefficiency relative to these imagined beings' doesn't cast doubt on our intelligence, why should the inefficiency of Marvin's brain relative to ours cast doubt on Marvin's intelligence? To suggest that it does would be arbitrary and *ad hoc*.[30]

Consider a Galton board (see Fig. 3). This device generates a bell-shaped pile of marbles at the bottom, when you feed marbles into it from the container at the top. It works, because between the top and bottom there are multiple layers of pegs. When a marble hits a peg, it has about an even chance of bouncing left or right. Since the number of paths (combinations of lefts and rights) that lead to a given one of the slots at the bottom is proportional to how close to the middle that slot is, a bell-shaped distribution of marbles results.

Now imagine trying to achieve the same bell-shaped distribution after removing all the pegs. To accomplish this, we install a mechanism that marbles feed into as they emerge from the reservoir at the top. This mechanism is a finely-calibrated variable marble shooter. Each time a marble enters it, it shoots the marble towards one or another of the receptacles at the bottom, with enough accuracy to put the marble in the targeted receptacle. The targeting mechanism adjusts the shooter's aim over the course of a given run, so that by the end of the run, it has shot marbles into receptacles so as to achieve the desired bell-shaped stack.

For the purpose of putting marbles into bell-shaped stacks, the Galton Board is much more efficient to build and operate than the Variable Marble

[30] The point isn't new: see (Block, 1981, 40) and (Braddon-Mitchell and Jackson, 2007, 90-91).

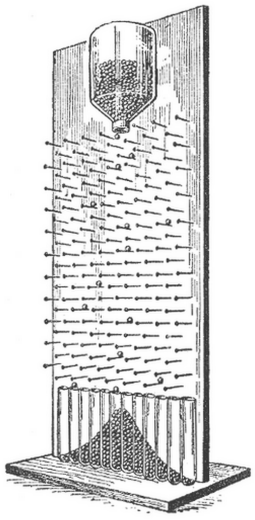

Figure 3: Galton board

Shooter, if only because it contains fewer moving parts (none, if you exclude the marbles). The Galton Board is to the Variable Marble Shooter as the human brain is to Marvin's Blockhead mechanism. Just as the Galton Board performs the same tasks as the Variable Marble Shooter using a more energy-efficient scheme that involves multiple layers of processing instead of a precisely calibrated targeting mechanism, the human brain performs the same tasks as the Blockhead system using a more energy-efficient scheme that involves multiple layers of processing instead of a precisely calibrated lookup table.

Is the Galton Board intuitively more intelligent (or "intelligent") than the Variable Marble Shooter? No. If your goal is to build a machine that sorts marbles into bell-shaped piles, then the intelligent choice is to build the Galton Board instead of the Variable Marble Shooter, assuming that you have limited resources to build your machine and operate it once built. But as far as the machines themselves are concerned, there's no intuitive sense in which the design of the Galton Board is more intelligent than that of the Variable Marble Shooter. This suggests that there is no deep connection between intelligence and efficiency.

7 Data compression

According to many psychoconservatives, a Blockhead mechanism like that which governs Marvin is too crude to sustain a mind. As Ned Block puts it, such a mechanism "lacks the kind of 'richness' of information processing requisite for intelligence."[31]

Conservatives attempt to spell out the relevant "richness" in various ways. One is in terms of polynomial time computation, considered above. Another is in terms of data compression. For example, according to Jens Kipper, a system's intelligence is proportional to the extent to which it relies on data compression.[32]

Data compression is a way to reduce the total size of the signals used to transmit information without reducing the amount of information transmitted.[33] The advantage of using data compression is that it decreases the cost of conveying information from one point to another: e.g., from a radio transmitter to a radio receiver, or from retinal surfaces to visual cortices, or from one cloud server to another. The basic idea is that since it costs less—i.e., consumes fewer resources—to transmit a shorter signal than it does to transmit a longer signal, we can increase a system's efficiency by choosing a system of representation in which smaller representations (e.g., in a binary system, shorter strings of `0`s and `1`s) represent pieces of information that need to be represented more often, reserving larger representations (e.g., longer strings of `0`s and `1`s) for pieces of information that need to be represented less often.

In order for this to work, the system doing the data compression has to be designed (or evolved) with some knowledge (or "knowledge") of the relative frequencies of the various pieces of information it's apt to be called upon to transmit. Data compression works only if the system plays the odds right, so that the information it's most frequently called upon to transmit is information that its encoding algorithm assigns the smallest representations.

---

[31] (Block, 1981, 28). He adds: "I wish I could say more about just what this sort of richness comes to."

[32] See Kipper (2019).

[33] This is true of lossless data compression. Lossy data compression reduces the sizes of signals without reducing the amount of information transmitted by more than a certain maximum.

The important point for present purposes is that data compression is a virtuous feature only when, and to the extent that, it allows a system to transmit information using fewer resources that it would have to use to transmit the same information without data compression (or with less compression). But we've already seen that efficiency is not necessary for intelligence. It follows that data compression is not necessary for intelligence.

We've considered polynomial time computation and data compression as forms of information processing allegedly required for true intelligence or cognition. These turn out not to bear the weight that psychoconservatives want to place on them. Might some other kind of information processing fit the conservative bill?

Here's a reason to think not.

Suppose we agree that some forms of computation are in some intuitive sense "richer" or more sophisticated than others. Then we can also agree that at some level of description, all of the computation that occurs in any physical system is highly *un*sophisticated. This is because whatever computations occur in a physical system supervene on the behavior of the system's constituent subatomic particles, which we can describe as natural computers that give various microphysical outputs (e.g., subatomic states or events) for various microphysical inputs, according to simple physical laws.[34]

Having components that perform such computations cannot be what makes the difference between an intelligent system like a human brain and an allegedly unintelligent system like Marvin's hydraulics, since such computations occur in both human brains and Marvin's hydraulics. So it seems that the conservative's claim must be that for a system to be intelligent, what's necessary is for the system to engage in sophisticated computation at *some* (but not necessarily every) level of description.

But then why can't it be at the level of the system's operational surface?

Insects, frogs, squirrels, and human beings all compete to survive, thrive, and propagate. Our brains all evolved to maximize the chances of replicating

[34] The same goes for any physical entity. As John Searle puts it, "From a mathematical point of view, anything whatever can be described as . . . instantiating or implementing a computer program. . . [T]he pen that is on the desk in front of me can be described as a digital computer. It just happens to have a very boring computer program. The program says: 'Stay there'." (Searle, 1984, 36)

the genes that generated them. How our brains achieve this (to the extent that they do) is largely determined by the outputs that their operational surfaces are disposed to give for various inputs. These input-output regimens are a form of computation: we can describe them in terms of algorithms for entering certain output states upon having been in certain input states.

Are they a rich or sophisticated form of computation? It's hard to say without hearing more about what "richness" and "sophistication" are supposed to involve, but in any intuitive, pre-theoretical sense of these terms, the surface-level computation that occurs in human brains is richer and more sophisticated than the computation that occurs in toasters, sponges, and the operational surfaces of frog and squirrel brains. The operational surface dynamics of human brains are certainly very complex, and typically very effective at solving the problems that face the organisms they govern; so, to the extent that complexity and effective problem-solving are signs of richness and sophistication, the computation that occurs in a human brain at a surface level of description is rich and sophisticated.

Psychoconservatives say that intelligence requires sophisticated computation at some level of description. Very well: Marvin satisfies this requirement in virtue of the sophistication of the computation that occurs in him at a level of description that captures the dynamics and dispositions of his brain's operational surfaces.

Conservatives can't object that internal-wiring computation takes precedence over surface computation, on the grounds that the internal-wiring computation underlies the surface computation. By that reasoning, the computation that occurs at the subatomic level takes precedence over that which occurs at the level of internal wiring, but, as already noted, subatomic-level computation is unsophisticated, and in any event ubiquitous.

Neither can conservatives insist that internal-wiring computation takes precedence over surface computation on the grounds that the former is more sophisticated than the latter. Consider someone whose individual neurons play the same roles as yours in relation to one another, but in whom each neuron uses far more complex algorithms than yours to compute outputs from inputs. (To make it vivid, we can imagine that little math professors play the role of individual neurons, calculating neuronal outputs for neuronal

inputs using needlessly sophisticated techniques.) Presumably conservatives don't want to say that this person is more intelligent than you.

If psychoconservatives want to claim that the only computational richness or sophistication that's relevant to a system's intelligence is richness or sophistication at a specific level of description distinct from surface-level description and sub-neuron-level description, they need to provide some justification for that claim. To insist without argument that a certain level of internal-wiring computation is the only computation directly relevant to intelligence would simply be to beg the question against the psycholiberal.

8 Psychodynamics

One psychoconservative reason for denying that Marvin has a mind is that there's nothing in Marvin corresponding to the psychological dynamics on display when, for example, a belief that p interacts with a belief that if p then q to give rise to a belief that q.[35] These relationships among our mental states—the kind involved in reasoning, deliberation, and decision-making, among other mental processes—are essential to our standing as rational, thinking beings. If such relationships don't occur in Marvin, he doesn't have a mind anything like ours.

However, there's nothing to prevent such relationships from occurring in Marvin. To convince you of this, I'll start by focusing on psychological states like beliefs, and assume a reductionist view of such states. Then I'll tell the same story again, but in nonreductionist terms that apply to all mental states, psychological as well as phenomenal.

If we look inside Marvin's hydraulics, we won't see anything that resembles sentences to play the role of beliefs, and we won't see anything that resembles sentence-use to play the role of inference, deliberation, etc. However, since the same is true if we look inside our own brains, this gives us no reason to think that Marvin differs from us mentally.

Many reductionist conservatives identify your beliefs with features of your brain that include your brain's internal wiring. These are what we may call

[35] This objection usually arises in discussions of the "language of thought" hypothesis, or the "systematicity of thinking"—see, e.g., Fodor (1987), Fodor and Pylyshyn (1988), and Davies (1992).

deep brain features. Unconscious or offline beliefs are deep brain features that instantiate unrealized or unactivated neural dispositions; conscious or occurrent beliefs are deep brain features that instantiate realized or activated neural dispositions. When your beliefs that p and that if-p-then-q lead you to form the belief that q, what's happening, according to psychoconservatives, is that a certain deep brain feature (your belief that p) combines with another deep brain feature (your belief that if p then q) to result in a third deep brain feature (your belief that q).

Reductionist liberals can say something exactly parallel to this: we just replace the conservative's deep brain features with operational surface features. Unconscious or offline beliefs are superficial brain features that instantiate unrealized or unactivated neural dispositions; conscious or occurrent beliefs are superficial brain features that instantiate realized or activated neural dispositions. When your beliefs that p and that if-p-then-q lead you to form the belief that q, what's happening, according to psycholiberals, is that a certain superficial brain feature (your belief that p) combines with another superficial brain feature (your belief that if p then q) to result in a third superficial brain feature (your belief that q).

In the liberal view, all the psychodynamical relationships are the same as in the conservative view. It's just that the relata are operational surface features, instead of deep brain features.

Similar remarks apply when it comes to nonreductionist liberalism and conservatism. According to nonreductionist conservatives, psychodynamical relations relate mental states that nomically supervene on corresponding deep brain features; according to nonreductionist liberals, psychodynamical relations relate mental states that nomically supervene on corresponding operational surface features. Either way, the states are capable of relating to each other in the ways characteristic of inference, deliberation, etc. So, whatever psychodynamics play out in you can also play out in Marvin.

Interactions among mental states (or their neural correlates) are not limited to the cognitive domain. There are also interactions among phenomenal states (or their neural correlates).

Take stereoscopic vision. Arguably, this involves visual experiences caused by the stimulation of one eye combining with visual experiences caused by the

stimulation of the other eye to give rise to distinct stereoscopic experiences.

Functionalists might explain this in terms of states of the internal wiring of the subject's brain. They might say that your brain is so configured that whenever the internal wiring that sustains your visual experience has the states and dispositions it has when your right eye is stimulated as well as the states and dispositions it has when your left eye is stimulated, it also has states and dispositions that sustain stereoscopic visual experience.

Psycholiberals can offer much the same explanation. We can say that your brain is so configured that whenever the operational surface features that sustain your visual experience have the states and dispositions they have when your right eye is stimulated as well as the states and dispositions they have when your left eye is stimulated, they also have states and dispositions that sustain stereoscopic visual experience. Once again, the relations can be the same on both accounts, only with different relata.

Similar remarks apply to other psychodynamical relations. What we believe often has a bearing on the phenomenal quality of our experience; this is perhaps most obvious in cases where a belief affects your mood, as when your belief that a forest fire is approaching your house gives you a feeling of anxiety. What we experience often has a bearing on what we believe; for example, it's partly because my visual experience has the phenomenal qualities it currently does that I believe I'm sitting in my living room.

A functionalist might explain these cases in terms of interactions or interdependencies among a subject's deep brain features. By this reckoning, your belief that your house is in jeopardy is a deep brain feature (state and/or disposition) that causes another deep brain feature, which is your feeling of anxiety; my current visual experience is a deep brain feature that causes another deep brain feature, which is my belief that I'm in my living room.

A psycholiberal can give the same explanation, in terms of operational surface features instead of deep brain features. Your belief that your house is in danger of burning down is a feature (state and/or disposition) of your brain's operational surface that causes another feature of your brain's operational surface, which is your feeling of anxiety; my current visual experience is a feature of my brain's operational surface that causes another feature of my brain's operational surface, which is my belief that I'm in my living room. The

liberal's account differs from the conservative's only in terms of relata, in a way that preserves the relevant psychodynamical relations.

9 Sources of conservative intuitions

Many people come to debates about AI and machine consciousness with strong conservative intuitions. I've argued that these intuitions receive no support from the considerations usually taken to vindicate them. So what is their real source?

One possible source of the intuition that Marvin is mindless is the operational transparency of the system that he has in lieu of a human brain. When you contemplate what goes on in Marvin's automated workbook, it seems incredible that *that* could give rise to thought, emotion, experience, etc.

But you know what else seems incredible? That a bunch of chemical reactions taking place in a blob of protein could give rise to thought, emotion, experience, etc. If we didn't know that our own minds arose from the blobs in our skulls, it would no more occur to us to that those blobs might sustain minds than that a plate of Jell-O might. It's just that our ignorance about the workings of our brains lets us imagine that if we only knew more about them, we'd see something mindlike in them that we don't see in Marvin's hydraulics.

There is also, I think, a related explanation for the intuition that Marvin and his ilk lack minds. It's the explanation David Braddon-Mitchell and Frank Jackson give for the intuition that the so-called China Brain lacks a mind.[36]

Some people think it's intuitively obvious that the China Brain doesn't sustain a mind. If they're right about this, it refutes not only the psycholiberal position that I favor, but the standard functionalist view that anything with the same network-wide functional organization as a normal human brain sustains a mind indistinguishable from that which the human brain sustains. According to Jackson and Braddon-Mitchell, however,

> the functionalist can reasonably deny the intuition. The source of the intuition that the system consisting of robot plus China brain lacks mental states like ours seems to be the fact that it would be so very much bigger than we are. We

[36] In the China Brain thought-experiment, billions of Chinese citizens send text messages back and forth in a way that perfectly parallels the patterns of synaptic signalling that take place in a normal human brain over some period of time; see (Block, 1980, 276-78).

> cannot imagine "seeing" it as a cohesive parcel of matter. We cannot see, that is to say, the forest for the trees. A highly intelligent microbe-sized being moving through our flesh and blood brains might have the same problem. It would see a whole mass of widely spaced entities interacting with each other in a way that made no sense to it, that formed no intelligible overall pattern from its perspective. The philosophers among these tiny beings might maintain with some vigour that there could be no intelligence here. All that is happening is an inscrutable sending back and forth of simple signals. They would be wrong. We think that the functionalist can fairly say that those who deny mentality in the China brain example are making the same mistake.[37]

A similar explanation applies to the intuition that Marvin lacks mental qualities like sentience and intelligence. At each time increment, the mechanism that serves as Marvin's brain performs a single, simple task: matching one alphanumeric string with another, moving a cursor from one column of a spreadsheet to another, closing a spreadsheet, opening a spreadsheet, etc. None of these tasks requires any intelligence, creativity, or ingenuity. If we just focus on them, we won't find any evidence of mental activity in Marvin's governing mechanism. But, as Jackson and Braddon-Mitchell point out, the same is true if we focus just on the simple tasks that the individual neurons of a human brain perform. These, too, require no intelligence, creativity, or ingenuity: after all, they're tasks that a single brain cell can perform. It would be a mistake to infer from this that my brain as a whole fails to sustain any mental activity; for the same reason, it's a mistake to infer from the mindlessness of the low-level processing that occurs in Marvin's brain that his brain as a whole fails to sustain any mental activity.

Curiously, Braddon-Mitchell and Jackson think that a Blockhead mechanism *does* fail to sustain any mental activity. They describe a version of Marvin whose brain consists of a vast decision tree, where each node of the tree receives input from the body, sends output to the body according to a pre-arranged program, and then simultaneously deactivates itself and activates a different node (also according to a pre-arranged program).[38] This version of Marvin does not differ from my version or Ned Block's original Blockhead in any important respect, but there are aesthetic differences, which perhaps explains

[37] (Braddon-Mitchell and Jackson, 2007, 109).

[38] See (Braddon-Mitchell and Jackson, 2007, 116-22).

why Jackson and Braddon-Mitchell fail to notice that their intuitions about Blockheads are subject to the same debunking as conservative intuitions about the China Brain. Each node of their Blockhead's decision tree performs a small number of simple, mindless tasks: receive encoded signal from body, search menu for matching code, activate next node. If we focus narrowly on these simple mindless tasks, we're apt to miss the forest for the trees, just like Jackson and Braddon-Mitchell's intelligent microbe.

The error that Jackson and Braddon-Mitchell identify (and commit) may arise from a more general tendency to conflate an agent's mental activities with the processing that underlies those activities. If Marvin's problem-solving methods were simply those employed by the internal wiring of his governing mechanism, then Marvin's method would be always to consult a vast compendium of spreadsheets. This is not the method of any human being. So, if you believe an agent's problem-solving methods are just those that underlie the input-output architecture of its governing mechanism, you'll conclude that Marvin is no human's mental equivalent.

But the belief is wrong. Marvin's problem-solving methods are not those that underlie the input-output architecture of his governing mechanism. Marvin doesn't consult a spreadsheet whenever he has to make a decision or react to an environmental change. He might go his entire life without using a spreadsheet; he might not even know what a spreadsheet is.

The use of spreadsheet algorithms is something that occurs in the internal wiring of Marvin's hydraulics, in a way that contributes to making him a competent problem solver. In the same way, the processes that occur in the internal wiring of a squirrel's brain contribute to making the squirrel a competent problem solver. It might be that rapid complex calculations occurring deep in a squirrel's brain play a crucial role in enabling the squirrel to solve the problem of navigating its way through the tree canopy leap by leap; this doesn't imply that such calculations are part of the squirrel's mental life. No more does the occurrence of spreadsheet manipulations deep in Marvin's brain imply that such manipulations are part of Marvin's mental life.[39]

[39] Why not say that the calculations that occur in the squirrel's brain are *unconscious* parts of the squirrel's mental life? Because a reasonable requirement for having an unconscious mental state is being capable of having the same or similar state consciously. That's why ten seconds ago you had the unconscious belief that gold is a precious metal, but did not

10 Conclusion

The current Zeitgeist in the philosophy of mind and AI is markedly conservative. In an extended whitepaper on machine consciousness released in August 2023, the co-authors—nineteen respected philosophers, psychologists, neuroscientists, and computer scientists—take psychoconservatism as a methodological axiom, dismissing the liberal viewpoint almost out of hand.[40]

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 large-language model similar to ChatGPT) was sentient. Here's a quote:[41]

> 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.

Lemoine is clearly taking a 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 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.

---

unconsciously perform the calculations that determined your neurons' firing patterns at that time, any more than you unconsciously performed the calculations that determined your hair follicles' rates of hair production.

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

[41] Blake Lemoine, quoted in Tiku (2022).

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 steeped in conservative thinking 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 superficially equivalent to ours have minds like ours.[42] From the liberal standpoint, 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.

The time may be closer than you think when we have to decide whether the artificial intelligences we build are genuinely intelligent, sentient agents. We would do well to prepare ourselves for this eventuality by reflecting now on whether the prevailing conservative requirements for sentience and intelligence are appropriate, or whether, as I have argued, they are excessively stringent. The cost of failing to do so could be very high.

[42](Block, 2002, 401-407).

References

Block, Ned. "Troubles with functionalism." In *Readings in Philosophy of Psychology, Volume I*, edited by Ned Block, Cambridge: Harvard University Press, 1980, 268–305.

———. "Psychologism and behaviorism." *Philosophical Review* 90, 1: (1981) 5–43.

———. "The harder problem of consciousness." *The Journal of Philosophy* 99, 8: (2002) 391–425.

Braddon-Mitchell, David, and Frank Jackson. *The Philosophy of Mind and Cognition*. Malden, MA: Blackwell, 2007, 2nd edition.

Burge, Tyler. "Individualism and the mental." *Midwest Studies in Philosophy* IV: (1977) 73–121.

Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon, and Rufin VanRullen. "Consciousness in artificial intelligence: insights from the science of consciousness." *arXiv* https://arxiv.org/abs/2308.08708.

Chalmers, David. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press, 1996.

Cobham, Alan. "The intrinsic computational difficulty of functions." In *Logic, Methodology and Philosophy of Science: Proceedings of the 1964 International Congress*, edited by Yehoshua Bar-Hillel, Amsterdam: North-Holland Publishing Company, 1965, 24–30.

Davies, Martin. "Aunty's own argument for the language of thought." In *Cognition, Semantics and Philosophy: Proceedings of the First International Colloquium on Cognitive Science*, edited by Jesús Ezquerro, and Jesús M. Larrazabal, Dordrecht: Springer, 1992, 235–271.

Dyson, Freeman. *Birds and Frogs: Selected Papers, 1990-2014*. Singapore: World Scientific, 2015.

Fodor, Jerry. *Psychosemantics*. Cambridge: MIT Press, 1987.

Fodor, Jerry A., and Zenon W. Pylyshyn. "Connectionism and cognitive architecture: a critical analysis." *Cognition* 28: (1988) 3–71.

Kipper, Jens. "Intuition, intelligence, data compression." *Synthese* 198, Suppl 27: (2019) 6469–6489.

Kirk, Robert. "Sentience and behavior." *Mind* 83, 329: (1974) 43–60.

Lewis, David. "Mad pain and Martian pain." In *Readings in Philosophy of Psychology, Vol. 1*, edited by Ned Block, Cambridge: Harvard University Press, 1980, 216–222.

Mäkelä, Jarmo. "Is reality analog or digital?" *https://arxiv.org/abs/1106.2541* .

Millikan, Ruth. *Language, Thought and Other Biological Categories*. Cambridge: MIT Press, 1984.

Newell, Allen, and Herbert A. Simon. "Computer science as empirical inquiry: symbols and search." *Communications of the ACM* 19, 3: (1979) 113–126.

Penrose, Roger. *The Emperor's New Mind*. Oxford: Oxford University Press, 1989.

Putnam, Hilary. "The meaning of "meaning"." *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science* 7: (1975) 131–193.

Searle, John R. *Minds Brains and Science*. Cambridge: Harvard University Press, 1984.

Smolin, Lee. *Three Roads to Quantum Gravity*. New York: Basic Books, 2001.

Tiku, Nitasha. "The Google engineer who thinks the company's AI has come to life." *The Washington Post,* June 11, 2022.