Synthese (2026) 207:259
https://doi.org/10.1007/s11229-026-05638-2

ORIGINAL RESEARCH

A liberal view of consciousness and cognition

Michael Pelczar[1]

Received: 12 May 2025 / Accepted: 6 May 2026
© The Author(s), under exclusive licence to Springer Nature B.V. 2026

Abstract
What does it take for an organ or artifact to sustain a conscious, intelligent mind? Must its detailed internal workings resemble those of an evolved human brain? Or is it enough if it has an input-output architecture comparable to that of a human brain, regardless of what internal structure supports the architecture? I argue that the answers to the last two questions are "no" and "yes," respectively. This sets the bar for machine sentience and intelligence much lower than most experts currently deem appropriate, and implies that we are closer than most people realize to building conscious, intelligent machines.

Keywords Consciousness · Cognition · Behaviorism · Functionalism · Artificial intelligence

1 Introduction

What does it take to have a conscious, intelligent mind? Recent technological advances give this question a new urgency. We are on the verge of building, if we have not already built, machines that satisfy some theorists' criteria for sentience and intelligence. 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 from the article:

> 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.

✉ Michael Pelczar
phimwp@nus.edu.sg

[1] Department of Philosophy, National University of Singapore, AS3, NUS, Singapore 117570, Singapore

Published online: 11 June 2026

> And I hear what they have to say, and that is how I decide what is and isn't a person.[1]

Two days after the publication of the *Post* exposé, Ned Block tweeted the following in response to Lemoine's comments:

> 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.

Unlike Lemoine, Block denies that we can infer that something has the mental characteristics of a person from the fact that it has the same kind of behavioral dispositions as a person. In Block's view, we can attribute mental properties only to beings that contain organs that process information using the same kind of electrochemical signalling that takes place in a naturally-evolved brain. Since LaMDA does not process information using that kind of electrochemical signalling, we should not, according to Block, believe that LaMDA has a mind.

Lemoine and Block stand at opposite ends of a spectrum of views on what it takes to build a verifiably conscious, intelligent mind with mental qualities comparable to our own (see Fig. 1).

At one end is behaviorism, the view that anything that has the same behavioral dispositions as a being with a mind has a mind indistinguishable from that being's. Realistically, having the right kind of dispositions requires having *some* natural or artificial brain that sustains the dispositions; but, according to behaviorists, it doesn't

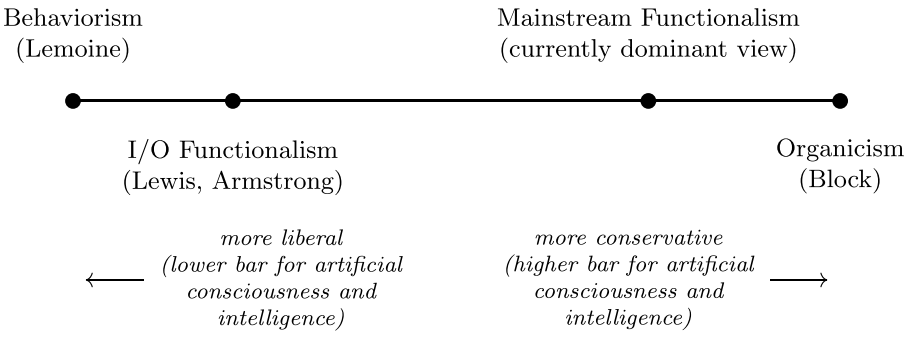


Fig. 1 A spectrum of views

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

matter at all what the mechanism is like, or how it sustains the dispositions: all that matters is that the dispositions themselves exist.[2]

At the opposite end of the spectrum is what I'll call organicism. According to organicists, in order to have a mind comparable to a human mind, you must be a biological organism. In this view, even if a robot's artificial brain implements the same algorithms and computational pathways as a natural human brain, it doesn't sustain a conscious, intelligent mental life, since it doesn't implement the algorithms and pathways using the biological processes that our evolved brains use.[3]

Neither of these views is very common today, and I won't discuss them further here. Instead, I want to focus on a pair of intermediate views, one closer to the organicist end of the spectrum, the other closer to the behaviorist end.

The view closer to organicism is what we may call *mainstream functionalism*. This is the widely-held view that when a human brain instantiates a mind, it does so in virtue of the brain's global functional organization, understood as the detailed neuron-level processing that takes place throughout the brain, including the processing that occurs in the brain's "internal wiring": the part of the brain that lies between the parts that receive direct input from and send direct output to the body that the brain governs. According to mainstream functionalists, when it comes to attributions of mind, internal wiring matters: an artificial brain might endow an attached body with dispositions identical to those that your brain endows your body with, but fail to instantiate any mental properties, due to having the wrong kind of internal wiring.

Mainstream functionalism is less liberal than behaviorism, since, as just noted, a humanoid robot might have the same bodily dispositions as a natural human being due to having an artificial brain that does not have the right kind of internal wiring. For example, the robot's brain might be an artificial neural network that converts sensory input to motor output using very different algorithms or network processes from those that a natural human brain uses; in that case, mainstream functionalism says that we should deny, or at least doubt, that the robot has a mind. At the same time, mainstream functionalism is more liberal than organicism, since an artificial brain could instantiate the same global functional organization as a human brain, despite being an inorganic artifact instead of a naturally evolved organ.[4]

The other intermediate view is what I'll call *input-output functionalism* or "I/O functionalism." This comes in several versions, but the basic idea is that the nature of a system's internal wiring is irrelevant to whether the system has a mind. According

[2] The classic sources for behaviorism are Wittgenstein (1958/2001) and Ryle (1949). The actual amount of brain tissue required to sustain normal human cognitive functions is surprisingly modest: see Feuillet et al. (2007).

[3] Block defends organicism most recently in Block (2026); see also Godfrey-Smith (2023) (though Godfrey-Smith doesn't fully commit to organicism). A view distinct from but related to organicism is that only biological material can implement the functional (structural and computational) features required for various mental properties: see Cao (2022). The difference between Block's view and Cao's is that according to Block, biology makes a non-functional contribution that is necessary for mental properties (or at least consciousness), while according to Cao, biology is necessary for consciousness only because it makes a functional contribution that non-biological phenomena can't make.

[4] Mainstream functionalism is widespread, but good examples include Putnam (1980), Lycan (1987), Chalmers (1996, pp. 246–75), Braddon-Mitchell & Jackson (2007, pp. 114–22), and (as regards consciousness in particular) Schwitzgebel (2025) and Schwitzgebel and Pober (2025).

to I/O functionalists, any system with the same input-output properties as your brain instantiates mental properties identical to those that your brain instantiates. Different versions of I/O functionalism differ in what they consider to be the relevant input-output properties, but, roughly, they're the properties a brain has in virtue of how the states of the brain's output surface (comprising the brain's motor neurons, which send output directly to the body's motor anatomy) depend on prior states of the brain's input surface (comprising the brain's sensory neurons, which receive input directly from the body's sensory surfaces).[5]

I/O functionalism is less conservative than mainstream functionalism, since it implies that if an artificial brain has the same I/O properties as your brain, it instantiates mental properties identical to yours, regardless of what algorithms, computations, or network processes occur in the brain's internal wiring. According to mainstream functionalists, we should deny, or at least doubt, that a humanoid robot with a next-generation GPT-powered brain has a mind, even if the robot's brain has the same I/O properties as your brain. According to I/O functionalists, such a robot has a mind indistinguishable from your own. At the same time, I/O functionalism is more conservative than behaviorism. Unlike behaviorists, who think that having the right kind of bodily dispositions is all it takes to have a mind, I/O functionalists think that having a mind requires having a brain that satisfies certain conditions, albeit conditions less stringent than those that mainstream functionalists impose.[6]

My goal in this paper is to argue for a version of I/O functionalism. This will differ from historic versions in two respects.

First, the version I defend does not say that there is a metaphysically or conceptually necessary connection between mental states and any kind of bodily behavior

[5] Neuroscientists sometimes call what I'm calling the brain's sensory neurons "sensory interneurons," and what I'm calling the brain's motor neurons "motor interneurons." This nomenclature is potentially confusing, since "interneuron" connotes a neuron that lies between the neurons that receive input to the brain and those that send output from the brain (what I'm calling the brain's sensory and motor neurons, respectively). The rationale for the scientific terminology is that there are neurons outside the brain that carry signals from sensory surfaces to the brain, and neurons outside the brain that carry signals from the brain to motor anatomy; neuroscientists also call these "sensory neurons" and "motor neurons," respectively. Since there seems to be no universally accepted nomenclature even within the neuroscience community, I am taking the liberty of adopting the nomenclature that is least confusing for the purposes of the present discussion.

[6] I/O functionalism isn't the only non-behaviorist theory situated left of mainstream functionalism. According to interpretivist theories of mental representation, your contentful mental states are those we must attribute to you in order to interpret your behavior as satisfying various norms (see, e.g., Williams (2019)); these theories are as liberal as behaviorism, at least when it comes to mental representation. In Cappelen and Dever (2025), Herman Cappelen and Josh Dever argue that existing LLMs are cognitive agents similar to ourselves (though Cappelen and Dever stop short of attributing consciousness to these agents). Though I don't think we're currently in a position to judge that existing LLMs are cognitive agents, the arguments that follow imply that we should at least consider the possibility that they are, and some of the considerations I put forward in Sect. 3 lend additional support to Cappelen and Dever's position. Apart from not extending their conclusion to phenomenal properties, the main difference between Cappelen and Dever's position and mine is that the former doesn't say which properties of LLMs are their cognitive properties: Cappelen and Dever's position is consistent with the behaviorist identification of cognitive properties with suitable behavioral/dispositional properties, as well as with identifications of cognitive properties with various types of I/O properties. In contrast, I identify our cognitive properties (and the cognitive properties of existing LLMs, if they have cognitive properties) with I/O properties of a specific type.

or behavioral disposition. In this, I differ from so-called analytic functionalists like David Lewis and David Armstrong, according to whom it is not just true but metaphysically necessary and even knowable *a priori* that a mental state is apt to cause certain types of bodily behavior and apt to sustain certain types of bodily dispositions. In my view, it is at most contingently true that there is this connection between mental states and bodily behavior and dispositions.

Second, when it comes to phenomenally conscious properties, my view is that these are not identical with or metaphysically supervenient upon I/O properties, but only nomically or naturally supervenient on I/O properties. In this, I also differ from Lewis and Armstrong, who hold that all mental properties, including phenomenal properties, are identical with I/O properties. I do agree with Lewis and Armstrong that *cognitive* properties are identical with I/O properties (though not with the behaviorally-individuated I/O properties with which Lewis and Armstrong identify cognitive properties).

I call my version of I/O functionalism "perimetric functionalism," and argue for it in two stages: first, I argue for the identification of cognitive properties with a certain type of I/O properties; then, I argue for the natural supervenience of phenomenal properties on I/O properties.

From here, the paper proceeds as follows. Section 2 characterizes mainstream and I/O functionalism in detail, by reference to the well-known Blockhead thought-experiment. Section 3 gives two arguments for reductionist I/O functionalism about cognition: a causal argument and an evolutionary argument. Section 4 reviews the classic I/O functionalist theories of David Lewis and David Armstrong, and identifies some serious problems with them, even when considered just as theories of cognition. Section 5 presents my preferred "perimetric" version of reductionist I/O functionalism about cognition. Section 6 addresses two important objections to perimetric functionalism. Section 7 argues for the natural supervenience of phenomenal properties on intrinsic I/O properties, using a version of David Chalmers's dancing qualia argument. Section 8 concludes.

2 Functionalism: mainstream vs. I/O

There's a part of you that governs your body, explaining why you engage in the bodily behavior and have the bodily dispositions that you do. Plausibly, this part is your central nervous system, or some sizable portion thereof. Following standard philosophical practice, I'll simply call it your *brain*. By your "body," I mean the part of you whose changes and dispositions are our usual (defeasible) basis for thinking that you have a mind with various mental properties. In this sense, your body is the part of you that typically provides third parties evidence of what, if anything, is happening in your mind. Plausibly, this part is your total organism minus your central nervous system, or some sizable portion of this total-organism-minus-central-nervous-system part.

Your brain has an *input surface,* consisting of the parts of it that the world outside your brain directly influences, and an *output surface,* consisting of the parts of your brain that directly influence the world outside. The locus of these influences is the

body to which your brain is attached; your body is, in this sense, your brain's immediate environment, and what we normally call "your environment" is the environment of that environment, i.e. the wider world that your body interacts with. In anatomical terms, a brain's input surface comprises the brain's sensory neurons, and the brain's output surface comprises motor neurons; in machine learning terms, the brain's input surface is its input layer, and the brain's output surface is its output layer.[7]

A brain's input and output surfaces together constitute what I'll call the brain's *operational surface*.

A brain's operational surface need not coincide with its anatomical surface. Nor does it depend for its existence on interactions with a body. Your brain's operational surface is the part of your brain that directly interacts with your body, but it is only a contingent truth about this part of your brain that it interacts with your body or anything else outside of your brain. (Analogously, a radio's antenna is the part of it that directly interacts with the radio's environment, but the antenna could exist even if it never interacted with the environment, perhaps for lack of anything else capable of sending or receiving radio signals.)

A brain's *I/O properties* are the properties it has in virtue of the states of its operational surface at various times, and of how states of its output surface depend on states of its input surface.

In addition to its input and output surfaces, a human brain includes a complex underlying network of brain cells that determines how changes in the input surface lead to changes in the output surface. This is what I'm calling the brain's *internal wiring*. In anatomical terms, a brain's internal wiring comprises the brain's interneurons (neurons of the brain other than sensory and motor neurons), together with pyramidal cells; in machine-learning terms, a brain's internal wiring comprises its "hidden layers." I'll call the properties of a brain's internal wiring *interstitial properties* of the brain.

According to I/O functionalists, any system with the same I/O properties as your brain has the same mental properties as your brain. According to mainstream functionalists, this is not the case: in their view, a system with the same I/O properties as your brain could fail to have any mental properties, due to its interstitial properties differing from your brain's in ways that result in the system's not having the kind of global functional organization that is (in their view) required for a mind. Here's an official statement of I/O and mainstream functionalism, in their most generic forms:

> **I/O Functionalism**: it is at least naturally necessary that systems with identical I/O properties have identical mental properties (if any), and that systems with similar I/O properties have correspondingly similar mental properties (if any).
>
> **Mainstream Functionalism**: it is at least naturally necessary that systems with identical global functional organization have identical mental properties (if any), and that systems with similar global functional organizations have cor-

[7] 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 input and output surfaces more exactly 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.

> respondingly similar mental properties (if any); however, it is naturally possible for there to be a mindless system with I/O properties identical to those of a system that has a mind.

The strongest form of I/O functionalism identifies all mental properties with I/O properties, and even the weaker version of I/O functionalism I defend here identifies cognitive properties with I/O properties. This is compatible with, but does not require, identifying cognitive *states* with I/O states (states a brain has just in case it instantiates appropriate I/O properties). An I/O functionalist might identify a subject's cognitive states with global functional states of the subject's brain; an I/O functionalist who makes this identification agrees with mainstream functionalists about the identities of cognitive states, but not about the identities of cognitive properties.

For practical purposes, it doesn't matter whether I/O functionalists identify cognitive states with I/O states or global functional states. Even I/O functionalists who identify cognitive states with global functional states will say that subjects whose brains have the same I/O properties have the same cognitive properties, regardless of whether the subjects' cognitive states are identical with the same types of global functional states. For example, they'll say that a robot whose brain is an artificial neural network with the same I/O properties as your brain has the property of believing that voles are mammals, if you have that property, even if the global states of the artificial brain that instantiate those properties differ greatly from the global states of your brain that instantiate them, due to differences in the brains' internal wiring.[8]

It will be useful to have an example of something that I/O functionalists say does, and mainstream functionalists say does not, have a mind. The literature helpfully provides such an example: the well-known Blockhead.[9]

Suppose we remove your brain from your skull, keeping it alive in a fluid-filled vat, and putting your body on life-support. For each place where a nerve fiber was cut to remove the brain, there is a pair of severed nerve-endings: one attached to the envatted brain, the other to the now brainless body. Some of the severed fibers carried messages from your body's sensory surfaces to your brain, others carried messages from your brain to various parts of your motor anatomy. To each pair of such severed endings, we attach a pair of wireless devices, each of which can send signals to or receive signals from the other. We configure the devices so that the connectivity between your brain and body is the same as it was before we removed your brain and put it into the vat.[10]

Now we remove the transmitters and receivers that are attached to your envatted brain, and integrate them with a computer that runs an automated Excel workbook

[8] For this reason, I'm not going to be too fastidious in my use of the terms "state" and "property," sometimes speaking of states where it would be strictly more appropriate (though wordier) to speak of properties. (I'll avoid such looseness of expression where it would be likely to cause confusion.).

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

[10] Daniel Dennett describes a scenario like this in (Dennett, 1981, pp. 310–23).

comprising a very large number of very long spreadsheets, numbered from Sheet 1 to Sheet i. Each sheet has three columns. The cells of the first column contain alphanumeric strings. Each string is uniquely associated with a possible state of your body's sensory surfaces (rods and cones, stereocilia, taste buds, Merkel cells, proprioceptors, etc). At each time increment, the computer running the workbook receives, via the receivers attached to the computer, a pattern of signals sent from the transmitters attached to the various elements of the body's sensory surfaces. This pattern gets translated into the string associated with the pattern in the computer's database. Using the standard Excel search function, 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 broadcasts, via the transmitters attached to it, the signal associated with S, which the receivers attached to the various elements of your motor anatomy receive, causing the body to change or remain the same in various ways (e.g., to engage in various motions, 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 composed in such a way that by replacing your natural brain with the system just described, your body engages in the same behavior, and has the same bodily dispositions, as it would if it were still governed by your original human brain.[11]

The being just described is a so-called Blockhead. The Blockhead's artificial brain, consisting of the computer running the Excel workbook and its attached receivers and transmitters, has the same I/O profile as your original human brain. The receivers constitute the artificial brain's input surface (counterpart of a natural brain's sensory cortex, consisting of the natural brain's sensory neurons), and the transmitters constitute its output surface (counterpart of a natural brain's motor cortex, consisting of the natural brain's motor neurons). The activation states of these surfaces over time mirror the activation states of the operational surface of your original, natural brain, and the dispositions of the artificial brain's output surface to be in certain states given prior states of the artificial brain's input surface mirror corresponding dispositions of the natural brain.[12]

Though mainstream functionalists disagree among themselves about what kind or degree of similarity to a human brain is enough to guarantee the presence of a human mind, they all agree that Blockhead's Excel-based brain does not clear the bar:

[11] Why use a workbook containing many spreadsheets, rather than a single sheet? Because if we use just a single sheet, then for each possible input I from the body, there is a certain output O such that the system outputs O whenever it receives an input of I; consequently, a being whose behavior was controlled by a single spreadsheet would be incapable of learning.

[12] As described, the thought-experiment ignores various complications, such as the possibility for the number and character (e.g., activation thresholds) of brain-body connections to change over time. We could embellish the thought-experiment to accommodate this possibility, but the unembellished thought-experiment doesn't oversimplify the synchronic details at any given moment, so it still serves to distinguish mainstream functionalists, who say there's no moment at which Blockhead has cognitive properties, from I/O functionalists, who say that at any given moment, Blockhead has the same cognitive properties as a human being whose brain has the I/O properties that Blockhead's brain has at that moment.

according to all mainstream functionalists, your Blockhead counterpart is a mindless machine.[13] In contrast, I/O functionalists think that your Blockhead counterpart *does* have a mind, and, in fact, a mind indistinguishable from your own.[14]

3 Arguments for I/O functionalism

In this section, I argue for identifying cognitive properties with I/O properties. By "cognitive properties," I mean general properties like *having beliefs, having desires,* and *being intelligent,* as well as more specific properties like *believing that there is a tiger in the vicinity, desiring not to be eaten by a tiger,* and *knowing how to elude tigers*. Cognitive properties do not include *phenomenal* properties or "qualia": the properties in virtue of which certain states are such that there's something it's like, subjectively, to be in them. Your cognitive properties are the mental properties you have in common with your hypothetical zombie twin, who duplicates you physically, but has no conscious experience; your phenomenal properties are the mental properties that you have but your zombie twin doesn't.[15]

Cognitive states are explanatory posits, albeit ones that we posit so instinctually that it's easy to forget they are posits. (Cognitive states differ from conscious states in this regard, which is why I argue for I/O functionalism about cognition separately from I/O functionalism about consciousness.) Here, I give two arguments for I/O functionalism about cognition, corresponding to the status of cognitive states as causal and teleological explanatory posits: causal, in that they explain the behaviors and dispositions they cause or sustain in the agents that have them, and teleological, in that they are the *raison d'être* of mechanisms with the relevant behavior-causing and disposition-sustaining powers. The first argument is reminiscent of the classic causal argument for analytical functionalism. The second is an evolutionary argument.

3.1 Causal argument

Define *mindlike behavior* as the kind of bodily behavior that we normally take as evidence for the presence of various mental features. Belieflike behavior is the sort of bodily behavior that we normally take as evidence that someone has beliefs; believing-that-voles-are-mammals-like behavior is the sort of bodily behavior that we normally take as evidence that someone believes that voles are mammals. Define

[13] Ned Block says that your Blockhead counterpart has the intelligence of a toaster (Block, 1981, p. 21); see also (Braddon-Mitchell & Jackson, 2007, p. 117) and (Chalmers, 1996, pp. 261–62).

[14] Though useful to differentiate mainstream from I/O functionalism, Blockheads are of little practical interest, since building a Blockhead brain would require more resources than the physical universe contains. More interesting are artificial neural networks that satisfy I/O but not mainstream functionalist criteria for consciousness and cognition; such networks may soon exist, if they do not already.

[15] If zombies are possible, there's a clear sense in which some are more intelligent than others, some but not others believe that the Earth is flat, some but not others like tofu, etc. If zombies are *not* possible, arguments similar to those for identifying cognitive properties with I/O properties support identifying phenomenal properties with I/O properties.

mindlike dispositions as the kind of bodily dispositions that we normally take as evidence that those who have them possess various cognitive features. For example, a desirelike disposition is the sort of bodily disposition that we normally take as evidence that the person who has the disposition has a certain desire.

Mindlike dispositions are more important that mindlike behavior, when it comes to the attribution of cognitive properties. An atheist playing a theist on stage, or pretending to be a theist to avoid persecution, might engage in the same mindlike bodily behavior as a theist, but he differs from a theist in his mindlike bodily dispositions; e.g., the atheist, but not the theist, is disposed to deny that God exists in off-stage contexts or more tolerant social settings. The importance of mindlike behavior is that it is defeasible evidence of mindlike dispositions, which in turn are evidence—also defeasible, but less easily defeated—of cognition. (When a brain state sustains a belieflike disposition, I'll say that the brain state "plays the belief role," and likewise for other mindlike dispositions.)

We can now state the following *causal argument for I/O functionalism:*

C1. Our cognitive features are those that are primarily responsible for our mindlike behavior and dispositions.
C2. The features that are primarily responsible for our mindlike behavior and dispositions are I/O features of our brains.
C3. Our cognitive features are I/O features of our brains. (C1, C2)

By saying that such-and-such features are "primarily responsible" for our mindlike behavior and dispositions, I mean that those features are typically the immediate causes of our mindlike behavior, and the direct sustainers of our mindlike bodily dispositions. C2 says that these features are I/O features of our brains. This is uncontroversial: with the exception of certain forms of reflex behavior, I/O features of our brains are primarily responsible for *all* of our behavior and dispositions.

C1 says that the features primarily responsible for our mindlike behavior and dispositions are our cognitive features: that, for example, our beliefs are typically the immediate causes of our belieflike behavior and the direct sustainers of our belieflike dispositions. According to so-called analytic functionalists, it is a necessary, *a priori* truth that the features primarily responsible for our mindlike behavior and dispositions are our cognitive features. The foregoing argument does not rely on any claim as strong as that. It's enough for present purposes that C1 is true: it doesn't have be a necessary or *a priori* truth.

Though neither necessary nor *a priori*, C1 is a platitude, or summation of platitudes, about cognition. For example, we ordinarily think of beliefs as states that typically cause belieflike behavior and sustain belieflike dispositions. Though beliefs do sometimes occur as effects or parallel concomitants of belieflike behavior and belieflike dispositions, we ordinarily, and, it seems, rightly, think that beliefs are typically causes and sustainers of belieflike behavior and dispositions, rather than effects or parallel accompaniments of them. Moreover, we ordinarily think of beliefs as causing and sustaining the relevant behavior and dispositions in a direct way, rather than merely by causing something that causes belieflike behavior, or sustaining something that sustains belieflike dispositions. There can be cases where a belief causes the rel-

evant behavior or sustains the relevant dispositions only indirectly, but we ordinarily, and, it seems, rightly, think of these as exceptions to the rule.

That said, one might challenge C1 on the grounds that our ordinary thinking is mistaken. Maybe beliefs, desires, etc. are not, as the platitude would have it, the primary causes and sustainers of mindlike behavior and dispositions, but something else, such as interstitial features of our brains that support those causes and sustainers, or a combination of such interstitial features and the I/O features they support. This challenge is only as strong as the reasons those who raise it provide for thinking that the "something else" is necessary for cognition. We return to this question below.

3.2 Evolutionary argument

Intelligence and other cognitive traits are evolved properties of our brains. To put it crudely, brains were evolved to think, just as lungs were evolved to breathe, and wings to generate lift. Their evolutionary *raison d'être* is to dispose the organisms they govern to behave in intelligent ways.

The dispositions that a brain confers on the body to which it's attached are a function of the brain's I/O properties. The I/O properties depend on interstitial properties of the brain's internal wiring, and natural selection also evolved those interstitial properties. But the evolutionary advantage of a given internal wiring scheme is primarily a matter of the I/O properties that the scheme supports. Evolution doesn't really care about the brain's internal wiring, except insofar as the internal wiring bears on the brain's surface-level I/O architecture, which is where the neural rubber hits the sensorimotor pavement.

We can think of this in terms of gene replication. How can the genes that build a brain improve their chances of being replicated? Primarily by building a brain that disposes the body to which the brain is attached to behave in ways that are conducive to the replication of those genes (along with other genes). Other desiderata are also relevant, such as the durability and energy efficiency of the brain that the genes build, but these desiderata are secondary to the primary desideratum of building a brain that disposes the body it governs to behave in selectionally advantageous ways. (If durability and energy efficiency were the main drivers of evolution, life wouldn't have evolved beyond microbial extremophiles.) Furthermore, all organ systems have their evolution constrained by generic desiderata like efficiency. If we want to know what drives the evolution of a particular type of organ such that it differs from other organs—e.g., brains versus lungs—we need to consider the distinctive functions of that type of organ. In the case of lungs, the distinctive functions are respiratory functions; in the case of brains, they are cognitive functions.

The only way that brain-building genes can build a brain that disposes the body to which it's attached to behave in ways that are conducive to the replication of those genes is by building a brain with suitable I/O properties, namely I/O properties that dispose the body to which the brain is attached to behave in advantageous ways. So, natural selection selects brain-building genes primarily for the I/O properties of the brains those genes build.

But it's also true that natural selection selects brain-building genes primarily for the cognitive traits of the brains those genes build: intelligence, memory, a capacity

for planning, etc. It follows that the cognitive traits of evolved brains are I/O traits. Call this the *evolutionary argument for I/O functionalism:*

E1. The primary evolved properties of our brains are cognitive properties.
E2. The primary evolved properties of our brains are properties in virtue of which our brains endow our bodies with advantageous behavioral dispositions.
E3. Our cognitive properties are properties of our brains in virtue of which our brains endow our bodies with advantageous behavioral dispositions. (E1, E2)
E4. The properties of our brains in virtue of which our brains endow our bodies with advantageous behavioral dispositions are our brains' input-output properties.
E5. Our cognitive properties are input-output properties of our brains. (E3, E4)

E4 should be uncontroversial. Our brains' I/O properties are precisely those that determine how our brains tell our bodies to behave in response to stimuli. A brain has the specific I/O properties it does largely due to the interstitial properties of the parts of the brain that connect the brain's input surface to its output surface, but these interstitial properties bear on the body's behavior only by bearing on the brain's I/O properties. It's true that a brain's I/O properties are a component of its global functional organization, and from this it follows that a brain's global functional organization fixes the bodily dispositions of the body it governs. But the global functional organization fixes those dispositions only in virtue of the I/O properties that the global organization includes: if the global organization had been different due to different internal wiring, but the I/O properties had been the same, the global organization would have endowed the body with the same behavioral dispositions.

One might challenge E2 on the grounds that our brains have important evolved properties unrelated to their behavior-governing properties. For example, the brain stem controls involuntary processes like breathing and heartbeat, and the hypothalamus regulates body temperature. To accommodate this, we can replace the first premise with, "The primary evolved non-autonomic properties of our brains are cognitive properties," and the second with, "The primary evolved non-autonomic properties of our brains are properties in virtue of which our brains endow our bodies with advantageous behavioral dispositions."

It's E1 that's most likely to encounter resistance. I've acknowledged that our brains possess evolved properties in addition to their primary behavior-governing properties. One might challenge E1 on the grounds that our cognitive properties might be, or essentially involve, one or more of these subsidiary properties; that is, one might contend that our brains' cognitive properties are not, or at least not just, their primary evolved properties. In particular, mainstream functionalists might say that the brain's cognitive properties are interstitial properties that underlie the brain's primary evolved properties (its behavior-guiding properties), or that the brain's cognitive properties are a combination of its primary evolved properties and various interstitial properties that underlie them.

In short, though our cognitive traits are evolved traits of our brains, there's still room for disagreement about exactly *which* evolved traits those are, and the fact that our brains' I/O traits are their primary evolved traits does not incontrovertibly favor identifying our cognitive traits with them, as opposed to evolved traits subsidiary to

the primary ones, or a combination of primary and subsidiary traits. Maybe, as mainstream functionalists contend, cognition has as much to do with *how* our brains perform their primary evolutionary function as it does with the fact *that* they perform it.

Like the main objection to the causal argument for I/O functionalism, this objection to the evolutionary argument succeeds only if mainstream functionalists can provide good reasons to think that certain interstitial properties are necessary for cognition (despite not being necessary for relevant I/O properties). As I now argue, all of the reasons they provide are inadequate.[16]

3.3 Challenges to the foregoing arguments

3.3.1 Adaptiveness and prior representation

According to David Braddon-Mitchell and Frank Jackson, in order to exhibit cognition, a system must have a certain kind of unscripted adaptiveness. In their view, it's because Blockhead's brain lacks this kind of adaptiveness that Blockhead is not a cognitive agent:

> Simple input-output devices exhibit massive causal dependencies between early and late stages. The state of a sundial or an amplifier or a carburettor that is responsible for its capacity to generate the appropriate outputs on Monday is typically a major causal factor in its capacity to do the job on Tuesday. The situation with much more complex structures like human beings is correspondingly more complex. How we respond to stimuli on Tuesday depends on all sorts of factors in addition to how we are on Monday, including what has impacted on us between the two days and what we have thought about in the interim. This is part of what confers on us the flexibility of response that makes us intelligent. Nevertheless, causal dependencies between earlier and later thoughts are crucial. It is just that how we respond in the future depends on a much more diverse range of factors than simply how we are in the past—what we have thought about and what has happened to us in the interim also enter the equation.
>
> The trouble with devices [like Blockhead's artificial brain] that work by look-up tree is that they lack the appropriate causal dependencies. The state that governs the responses to inputs early on plays the wrong kind of role in causing the state that governs the responses later on. This is because, for the most part,

[16] The non-I/O properties that people have suggested are necessary for cognition include more than those that I consider below. I limit myself to the non-I/O properties that, in my experience, people are most inclined to think are necessary for cognition. Block (1981) and (Braddon-Mitchell & Jackson, 2007, pp. 107–28) cover various properties that I omit here, arguing, to my mind convincingly, that none of them is really necessary for cognition (these properties include analogicity of information processing, causal indeterminacy, and operational robustness or redundancy). Bailey (2016) argues that whether a system is a cognitive agent depends on whether its parts compose a single entity, in the sense germane to the special composition question of metaphysics; be that as it may, systems like Blockhead appear to satisfy the relevant composition requirement no less clearly than human brains.

> the Blockhead is static. It is mostly written down in advance, and the only thing that varies is which node is active.[17]

Both Blockhead's brain and a human brain change over time by receiving different input from a body's sensory surfaces from moment to moment, and sending appropriate output to the body's motor anatomy. In addition, a human brain undergoes interstitial changes that result in changes in the dependencies between the brain's input and output surfaces. That these dependencies do change is obvious, and in any event necessary for the brain to learn, or gain or lose any cognitive features (e.g., acquire a new desire, or lose an old belief). Earlier states of a human brain play a crucial role in determining what output the brain is apt to give at later stages in response to input received at those stages—a role they play by affecting the brain's interstitial properties. The human brain is in this sense *adaptive,* unlike, say, a typewriter, which always gives the same output for the same input.

According to Braddon-Mitchell and Jackson, this is not true of Blockhead's brain. Why not? Because Blockhead's brain is "static," in the sense that all the various ways that it might translate input to output at a given moment are written down in advance (in the form of all those spreadsheets); the only thing that varies is which spreadsheet is active. According to Braddon-Mitchell and Jackson, this implies that Blockhead does not have a mind.

There are two things to say about this argument.

First, from the fact that Blockhead's brain is static in the relevant sense, it doesn't follow that its later I/O dispositions don't depend on earlier states of the brain. They very much do, since the Blockhead brain's I/O dispositions at a given moment depend on which spreadsheet is active at that moment, and which spreadsheet is active at any given moment depends on which spreadsheets were active at earlier moments together with what input the Blockhead brain received from Blockhead's sensory surfaces. Blockhead's brain is adaptive; it's just that its adaptiveness is due to its having an operational architecture very different from that of a human brain.

The only potentially important difference between Blockhead's brain and human brains that Braddon-Mitchell and Jackson have really identified is that all of the possible synchronic I/O dispositions of Blockhead's brain—all input-output dependencies that might characterize the brain at a given stage—are represented in advance, whereas this is not true of a human brain.

But why does this matter? Suppose that a Laplacean Demon wrote down all of the possible synchronic I/O dispositions of my (human) brain before I was born. Just as each stage of Blockhead's brain is such that there is a pre-existing representation of it (in the form of one of the spreadsheets in the automated workbook), each stage of my brain is such that there is a pre-existing representation of it (in the form of one of the pages in the Demon's notebook). This obviously would not prevent me from having a mind.

[17] (Braddon-Mitchell & Jackson, 2007, p. 120). By a "look-up tree," Braddon-Mitchell and Jackson mean a setup like the Excel workbook described earlier. A "node" of a look-up tree is just a spreadsheet in the workbook, and the node that is "active" at a given time is the sheet that receives input from Blockhead's sensory surfaces at that time and then sends output to Blockhead's motor anatomy.

Braddon-Mitchell and Jackson might point out that, unlike the spreadsheets in the automated workbook, the pages of the Demon's notebook play no role in what state my brain is in at any given moment. That's true, but, again, irrelevant. We can imagine that the Demon's notebook *does* play a role in determining what state my brain is in at any given moment. We can imagine that the state of my brain at one moment is intrinsically powerless to cause my brain to be in a different state at the next moment, and that the changes in my brain over time are due to a continual series of assists from the Demon, who, at each moment, causes my brain to enter the state described on one of the pages of his notebook, depending on which states my brain was in at previous moments, and what input it receives at the present moment. (According to many early modern philosophers, this is how all diachronic causation works, with God playing the Demon's role.) In this scenario, there is no appreciable difference between my brain and Blockhead's, in terms of how later I/O states depend on earlier I/O states and system inputs. But it's a scenario in which I have exactly the cognitive properties that I actually have. It can hardly be that I'm deprived of beliefs, desires, etc., by the fact that earlier states of my brain cause later states via the intermediation of a third party; a belief in the early modern theory of diachronic causation might be false, but it's not self-refuting. I conclude that there is no cogent objection to I/O functionalism, here.

3.3.2 Reasoning

One requirement for cognition is that the cognitive system exhibit an appropriate "systematicity of thinking" of the sort that's on display when, for example, an agent infers that q from the premises that p, and that if-p-then-q.[18] This kind of cognitive activity is essential to our standing as rational, thinking beings. If it doesn't occur in Blockhead, he doesn't have a mind anything like ours.

If we look inside Blockhead's artificial brain, 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 Blockhead differs from us mentally.

If Blockhead fails to be a reasoner, it's presumably due to his not having states that relate to one another the way that beliefs relate in us when we reason. The relevant relation is one of dependence. Suppose I infer (1) that the Sun is about to rise, from the premises (2) that the rooster is crowing and (3) that if the rooster is crowing, then the Sun is about to rise. In order for this to be an inference, rather than a haphazard sequence of thoughts, my having the belief that the Sun is about to rise must depend on my having the other two beliefs. The existence of such a dependence is what makes the difference between my inferring that the Sun is about to rise from the stated premises, and serially thinking, "the rooster is crowing," "if the rooster is crowing, then the Sun is about to rise," and "the Sun is about to rise" without making any inference.

An agent can be a reasoner without being a good reasoner. Someone might infer (a) that the Sun is not about to rise, from the premises (b) that the rooster is not crow-

[18] See, e.g., Fodor (1987), Fodor and Pylyshyn (1988), and Davies (1992).

ing, and (c) that if the rooster is crowing, the Sun is about to rise. This inference is invalid, but it's still an example of reasoning, provided that the belief that (a) depends on the beliefs that (b) and that (c). In contrast, someone who believes (a), (b), and (c), but whose belief that (a) does not depend on his beliefs that (b) and that (c), cannot be accused of invalid reasoning.

According to mainstream functionalists, having a belief requires more than instantiating (or having a brain that instantiates) certain I/O properties. It requires having brain states with certain interstitial properties, perhaps together with certain I/O properties; call these "deep-brain states." You believe that p by having a brain in a certain deep-brain state D_1, you believe that if-p-then-q by having a brain in a certain deep-brain state D_2, and you believe that q by having a brain in a certain deep-brain state D_3. You infer that q from the premises that p and that if-p-then-q only if D_3 depends for its existence on the existence of D_1 and D_2.

Perimetric functionalists can say something exactly parallel to this: we just replace the mainstream functionalist's deep-brain states with I/O states. The parallel holds, even if we identify I/O states with instantiations of I/O properties, rather than with the global functional states that instantiate such properties. You believe that p by having a brain in a certain I/O state S_1, you believe that if-p-then-q by having a brain in a certain I/O state S_2, and you believe that q by having a brain in a certain I/O state S_3. You infer that q from the premises that p and that if-p-then-q only if S_3 depends for its existence on the existence of S_1 and S_2. Since Blockhead's brain has the same I/O states as yours, standing in the same relations of dependence as your I/O states, he is as much a reasoner as you are, by this account.

The foregoing remarks gloss over some important complications. Not just any old dependence of one belief on others suffices for a relation of inference. If I form the beliefs that (2) and that (3) and this somehow causes me to have a stroke, I don't count as inferring that (1), even if the stroke damages my brain in a way that happens to cause me to believe that (1). Saying exactly what kind of dependence relation beliefs must stand in for there to be an inference is beyond the scope of this paper. For present purposes, the details don't matter, since cases like the stroke complicate accounts of inference regardless of whether the relata of inferential relations are I/O states or deep-brain states, and the most promising ways to deal with such complications are equally available to mainstream and I/O functionalists.

For example, maybe "at time t, x infers *q* from *p* and *if-p-then-q*" means that in all nearby possible worlds where x believes that p and that if-p-then-q, x believes that q.[19] Arguably, the stroke victim's beliefs do not satisfy this condition, since there are nearby possible worlds where the stroke does not scar my brain with a belief that the Sun is about to rise. But this is true whether beliefs are I/O states or deep-brain states.

Alternatively, maybe "x infers *q* from *p* and *if-p-then-q*" means that the probability of x believing that q, conditional on his believing both that p and that if-p-then-q, is high (i.e., above some suitable threshold). The stroke victim's beliefs arguably do not satisfy this condition, since the stroke's causing a belief that the Sun was about to rise was a low probability event, even given my prior beliefs that the rooster was crowing

[19] This is reminiscent of the so-called safety condition for inferential knowledge: see (Sosa, 1999, p. 146), (Williamson, 2000, p. 147), and (Pritchard, 2007, p. 292).

and that if the rooster crows, the Sun is about to rise. But again, this is true whether beliefs are I/O states or deep-brain states.

Perhaps there is a plausible theory of inferring that handles deviant cases like the stroke in a way that works only on the assumption that cognitive states like beliefs are deep-brain states. I am not aware of such a theory, but even if one were to emerge, the implications for I/O functionalism would not be dire. As noted earlier, while all I/O functionalists identify cognitive *properties,* such as the property of believing that p, with I/O properties, it's open to an I/O functionalist to identify cognitive *states,* such as a belief that p, with global functional states of the brain. In this version of I/O functionalism, beliefs are global functional states that have both I/O and interstitial properties, though what makes them beliefs (rather than cognitive states of some other sort, or non-cognitive states) is their I/O properties. An I/O functionalist who takes this view can give an account of inference and related cognitive activities that is effectively identical to a mainstream functionalist account, since mainstream functionalists also identify beliefs with global functional states: it's just that in the mainstream view, what makes the states beliefs isn't (just) their I/O properties.

3.3.3 Efficiency

Some mainstream functionalists suggest that in order to count as a cognitive agent, one's brain must exhibit an appropriate kind of efficiency. In their view, Blockhead is not a cognitive agent, despite having the same I/O properties as an ordinary human brain, because his brain relies on a grossly inefficient search strategy.[20]

For example, according to Allen Newell and Herbert Simon, "the task of intelligence . . . is to avert the ever-present threat of the exponential explosion of search."[21] By "exponential explosion of search," Newell and Simon are alluding to so-called explosive algorithms that have to be applied n^i times to give an output for an input of size i (where i and n are integers and $n > 1$). For example, suppose algorithm A takes strings of `0`s and `1`s as inputs, and gives other such strings as outputs. A is "explosive" if, for all i, it takes 2^i applications of A for A to deliver an output, given a string i digits long as input.

Blockhead's brain does rely on an explosive search strategy, so if Newell and Simon are right, Blockhead is not intelligent.[22]

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

[20] (Block, 1981, p, 38) attributes this suggestion to Daniel Dennett; see also the comments by Allen Newell and Herbert Simon quoted in the next paragraph.

[21] (Newell & Simon, 1976, p. 123), quoted in (Block, 1981, p. 38). Newell and Simon also make the more general claim that intelligence is essentially the avoidance of prohibitively costly problem-solving strategies (Newell & Simon, 1976, p. 121); this is implausible, since it implies that intelligence couldn't exist in an environment free from resource constraints.

[22] The number of spreadsheets that Blockhead's brain requires to 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, the algorithm has to choose from among n^i spreadsheets to open next, since there are n^i possible sequences of sensory inputs requiring an appropriate motor response (where n is the number of possible total states of the input surface of Blockhead's brain=the number of possible overall activation states of Blockhead's sensory organs).

perform each iteration of algorithm A, it would take the computer about 30 billion years to give an output for a 100-digit input.

Functions describing algorithms that are *not* impractical or unfeasible in this way are known as functions "computable in polynomial time." 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 quantity 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 nanosecond when we run it on the picosecond computer.[23]

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.

The argument for the claim that using polynomially executable algorithms is necessary for intelligence therefore boils down to an appeal to efficiency. But how does the relatively low efficiency of Blockhead's brain suggest that Blockhead lacks a mind? Why doesn't it just show that Blockhead has a mind sustained by a relatively inefficient brain?

Ectotherms use less energy than endotherms to maintain viable body temperature. That doesn't mean that endotherms don't thermoregulate: it just means that they thermoregulate by less energy-efficient means than ectotherms. Likewise, the fact that humans use less energy than Blockheads to think doesn't entail that Blockheads don't think: maybe they just think by less energy-efficient means.

We can imagine a world whose natural laws differ from our world's in such a way that it takes much more energy to power a system with the structure of a human brain than one with the structure of a Blockhead brain, but in which usable energy is so plentiful that there are human brains that function just like ours. Such a world isn't one where human brains fail to sustain minds. It's a world where human brains sustain minds by consuming more energy than they do in our world. (Analogously, there are possible worlds where it takes more energy to power light bulbs than in ours, but in which tungsten filaments inside evacuated glass enclosures incandesce due to being provided with the requisite amount of energy; these aren't worlds in which light bulbs fail to provide illumination: they're worlds where light bulbs provide illumination by consuming more energy than they consume in our world.)

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

In general, it's hard to see how a system's efficiency bears on whether the system has 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 required for cognitive properties, such as intelligence. Even though our brains are more efficient than Blockhead's, they are not the most efficient conceivable brains capable of performing the functions they do. Evolution finds efficient solutions, but not, in general, the most efficient possible solutions. We can imagine beings who have cognitive capacities identical to ours, in virtue of having brains even more efficient than ours. It would be a mistake to say that such beings lack intelligence because their brains are more efficient than ours. It would also be a mistake 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 Blockhead's brain relative to ours cast doubt on Blockhead's intelligence?[24]

3.3.4 Data compression

According to Jens Kipper, a system's intelligence is proportional to the extent to which it relies on data compression. If this is correct, then Blockhead is not an intelligent agent, since no data compression takes place in his brain.[25]

Data compression is a way to reduce the total size of the signals used to transmit information without reducing the amount of information transmitted.[26] 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.

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 need to transmit the same information without data compression (or with less compression). But we've already seen that efficiency

[24] See also (Block, 1981, p. 40) and (Braddon-Mitchell & Jackson, 2007, pp. 90–91).

[25] See Kipper (2019).

[26] 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.

is not necessary for intelligence. It follows that data compression is not necessary for intelligence either.

4 Traditional I/O functionalism

I've argued that our cognitive properties are I/O properties of our brains. But which I/O properties, exactly? We can describe the operational surface of a brain at different levels of generality, i.e. in terms of more or less generic input-output properties. We can describe the operational surface of a brain purely in terms of its intrinsic properties, or in terms that include extrinsic properties that the surface has in virtue of its relationships to other things. Different versions of I/O functionalism result from identifying our cognitive properties with different types of I/O properties: e.g., more versus less generic, or intrinsic versus extrinsic.

We can distinguish four main types I/O functionalism: *neo-behaviorist functionalism,* which identifies a subject's cognitive properties with I/O properties of the subject's brain individuated by the mindlike dispositions they actually sustain; *black box functionalism,* which identifies a subject's cognitive properties with I/O properties of the subject's brain individuated by the mindlike dispositions they have the power to sustain; *Lewisian functionalism,* which identifies a subject's cognitive properties with I/O properties of the subject's brain individuated by the mindlike dispositions they actually sustain in normal members of the population to which the subject belongs; and *perimetric functionalism,* which identifies a subject's cognitive properties with I/O properties that the subject's brain has intrinsically, and not just by having the power to sustain mindlike bodily dispositions. In this section, I discuss the first three types of I/O functionalism.

4.1 Neo-behaviorist functionalism

Neo-behaviorist functionalism identifies cognitive states with states that actually sustain relevant mindlike dispositions. This is different from behaviorism, which identifies cognitive states with the mindlike dispositions themselves, but it is hardly more plausible than behaviorism (which probably explains why nobody has ever endorsed neo-behaviorist functionalism, as far as I am aware). If we identify cognitive states with states that actually sustain relevant mindlike dispositions, we have to deny that paralytics, victims of locked-in syndrome, and the fabled brain-in-a-vat have beliefs, desires, or any other cognitive states, since they do not have any bodily dispositions. Even if a brain is completely indistinguishable from yours in all intrinsic respects, neo-behaviorist functionalism implies that it fails to instantiate any cognitive properties, if, like a BIV or the brain of a total paralytic, it fails to sustain any mindlike dispositions.

4.2 Black box functionalism

To avoid denying that paralytics etc. are cognitive agents, we might identify cognitive properties with properties that endow what has them with the *power* to sustain

relevant mindlike dispositions; e.g., we might identify beliefs with states that have the power to sustain belieflike dispositions. This is what David Armstrong does:[27]

> The concept of a mental state is the concept of that, whatever it may turn out to be, which is brought about in a man by certain stimuli and which in turn brings about certain responses.[28]

> [I]n talking about mental states we are simply talking about states of the person apt for the bringing about of behaviour of a certain sort.[29]

Armstrong doesn't explicitly identify a subject's cognitive properties with I/O properties of the subject's brain, but this identification is implicit in his view, since a brain has its behavior-causing powers in virtue of its I/O properties. If two subjects have brains with states that are identical in terms of what stimuli are apt to cause them and what behavior they are apt to cause—i.e., states with identical I/O properties—the subjects are cognitively indistinguishable, in Armstrong's view, even if their brains have very different internal wiring.

Since an envatted brain or the brain of a paralytic or lock-in has the power to cause mindlike behavior and sustain mindlike bodily dispositions, such individuals can have beliefs and other cognitive states, according to black box functionalism. Black box functionalism therefore avoids the most obvious pitfalls of behaviorism and neo-behaviorist functionalism.

The problem is that brains have too many powers like this. Suppose Joe is sane, and that Joe's brain sustains sanity-like bodily dispositions (in Joe's body). We temporarily disconnect Joe's brain from his body, and then reconnect it in a new way, such that Joe's brain—which has not undergone any relevant changes in the interim—no longer sustains sanity-like bodily dispositions, but does sustain insanity-like dispositions. Since Joe's brain remains unchanged throughout this process, and since he was not insane at the outset, we have compelling reason to think that he is not insane after the brain-body reconnection procedure. Yet, Joe's brain had the *power* to sustain insanity-like bodily dispositions all along, as demonstrated by the fact that it actually sustains such dispositions after the brain-body reconnection.[30]

This poses a problem for black box functionalism. Any brain that has the power to sustain one suite of mindlike bodily dispositions has the power to sustain any number of alternative suites of mindlike bodily dispositions. Your brain currently has the power to sustain not only the mindlike bodily dispositions that it actually sustains right now, but mindlike dispositions unlike any that you have ever had or ever will have. Black box functionalism implies, implausibly, that you have cognitive states corresponding to all of these unactualized mindlike dispositions, since your brain has the power to sustain them.

[27] See also Place (1956) and Smart (1959).

[28] (Armstrong, 1968, p. 79).

[29] (Armstrong, 1968, p. 89).

[30] The origin of cases like this is Lewis (1980).

4.3 Lewisian functionalism

Neo-behaviorist functionalism implies that a total paralytic with a brain intrinsically indistinguishable from yours has no cognitive properties; black box functionalism implies that you have many cognitive properties that you do not, in fact, have.

To avoid these implications, David Lewis identifies a subject's cognitive properties with the properties of the subject's brain that normally sustain relevant mindlike dispositions in members of the population to which the subject belongs. The properties of my brain that sustain my belieflike dispositions—i.e., play the belief role—are I/O properties of my brain that normally play the belief role in (other) members of the human species; call these ϕ-properties. A total paralytic whose brain has ϕ-properties has beliefs, according to Lewis, since he has the properties that normally play the belief role in members of his population, even though they don't play the belief role in the paralytic himself. Both before and after the reconfiguration of Joe's brain-body connection, Joe's brain has I/O properties with the power to sustain insanity-like dispositions. However, at no point does Joe's brain have the I/O properties that normally sustain insanity-like dispositions in members of Joe's population (the human species), so, according to Lewis, Joe is never insane, which is intuitively the correct verdict.[31]

Lewis's theory delivers the right verdicts on these cases by identifying the cognitive states of a given subject with brain states of that subject individuated partly by their relationships to other subjects' brain states. In Lewis's view, cognitive properties are not intrinsic properties of the individuals who have them, but relational properties that a subject has partly in virtue of how his brain states relate to the brain states of other members of his population.

Construing cognitive properties as relational properties in this way allows Lewis to avoid the pitfalls of neo-behaviorism and black box functionalism, but it gives rise to other, equally serious problems.

As Lewis acknowledges, there could be a person with a brain intrinsically indistinguishable from mine, but connected to his body in such a way that it sustains mindlike dispositions very different from my own (as in Joe's case); perhaps, unlike me, he doesn't have believing-that-voles-are-mammals-like dispositions, and does have believing-that-voles-are-lizards-like dispositions. Furthermore, as Lewis also acknowledges, such a person could come into existence by a highly improbable, though nomically possible, spontaneous assemblage of atoms.

The spontaneously generated Joe-like version of me is the only member of his population. The brain states sustaining his belief-that-voles-are-lizards-like dispositions are therefore those that normally sustain such dispositions in members of his population. Lewis's theory implies that he believes that voles are lizards, despite having a brain intrinsically indistinguishable from my actual brain.

There could also be a spontaneously generated person with a brain intrinsically indistinguishable from mine that sustains no bodily dispositions, due to total paraly-

[31] See Lewis (1980). There is actually a precursor to an account like Lewis's in Armstrong: see (Armstrong, 1968, pp. 142–44), where in the course of giving his analysis of having a goal or purpose, Armstrong appeals to a brain state's effects "in normal circumstances."

sis, locked-in syndrome, or envatment. Lewis's theory implies that such a person has no cognitive states, since none of his brain states sustains mindlike dispositions, which, in the nature of the case, means that none of his brain states normally sustains mindlike dispositions in members of his population.

We can also imagine a population in which people have naturally evolved in such a way that brain states that play the belief role in some of them play the desire role in others; which brain states play the belief or desire role in a given member of the population depends on his or her genetics, but children don't always take after their parents in this regard. Assuming that the population is evenly divided between these two types, there is no such thing as *the* brain state that normally plays the belief role in members of the population. Lewis's theory implies, implausibly, that no one in this population has beliefs.[32]

The main objection to neo-behaviorist functionalism was its implication that a total paralytic has no cognitive properties, even if his brain is intrinsically indistinguishable from yours. But Lewisian functionalism has essentially the same implication: given a spontaneously generated total paralytic with a brain intrinsically indistinguishable from yours, Lewisian functionalism implies that the paralytic has no cognitive states, since he has no brain states that normally sustain relevant mindlike dispositions in members of his population (i.e., in himself). If you're willing to accept this, it's hard to see why you would balk at the suggestion that a non-spontaneously-generated paralytic lacks cognitive states, and therefore hard to see why you would find neo-behaviorist functionalism objectionable. To put it the other way around: if you are *not* willing to accept neo-behaviorist functionalism, you shouldn't be willing to accept Lewisian functionalism either.

5 Perimetric functionalism

The problems with Lewis's theory arise from his identification of a subject's cognitive properties with relational properties of the subject's brain: properties that the subject's brain has in virtue of possessing I/O properties that normally play certain roles in members of the subject's population. To avoid these problems, we need to identify a subject's cognitive properties with properties that the subject's brain has in virtue of possessing I/O properties that it possesses independently of how the subject's brain relates to anything else.

A human brain's input surface comprises a finite number of sensory neurons, and its output surface comprises a finite number motor neurons. The motor neurons are examples of what I'll call *output elements,* and the sensory neurons examples of what I'll call *input elements*.

Sensory and motor neurons are just special kinds of input and output elements: if we replace each of your sensory and motor neurons with a functionally equivalent synthetic micro-prosthesis, then your brain no longer has sensory or motor neurons, but it still has input and output elements.

[32] Or, as Lewis would put it, that there is "no determinate fact of the matter" about whether any given member of the described population has beliefs (or desires): (Lewis, 1980, p. 220).

Perimetric functionalism about cognition identifies a brain's cognitive properties with intrinsic input-output properties of that brain. Specifically, it identifies a brain's cognitive properties with properties that the brain has in virtue of its input and output elements being in various activation states at various times, together with dispositions for the brain's output elements to be in certain states, given earlier states of the brain's input elements (more on these dispositions in a moment). I call these the brain's *perimetric properties*. They are intrinsic to the brain, in the standard sense that the brain could have them, even if it were the only thing that existed.[33]

A brain's perimetric properties include both the properties it has in virtue of the activation states of its input and output elements, and the dispositions for its output elements to be in certain activation states, given prior activation states of its input elements. Let me say something more about the dispositions.

At any given time, your brain satisfies numerous conditionals of the form, "If your brain's input surface is now in state S_i, its output surface will be in state S_o at the next moment (assuming that your brain survives to the next moment)." We can understand these as counterfactual conditionals of the form:

> If your brain's input surface were now in S_i, its output surface would be in S_o at the next moment (assuming that your brain were to survive).

—with the caveat that some of the conditionals have true antecedents. Alternatively (and, in my opinion, preferably), we can say that at any given time your brain satisfies numerous conditional probabilities of the form:

> Pr(your brain's output surface is in S_o a moment from now | your brain survives and its input surface is now in S_i)$=x$.

—where x has some appropriately high value. Your brain's *I/O dispositions* are the dispositions it has in virtue of satisfying various conditionals or conditional probabilities like these at various times. These dispositions are intrinsic properties of your brain; indeed, they are even intrinsic properties of your brain's operational surface, since there are metaphysically possible worlds where your brain's output surface depends in the relevant ways on your brain's input surface despite those surfaces being the only things that exist.

In the actual world, your brain has the I/O dispositions it does largely because of its interstitial properties: the internal parts of the brain that receive input from the brain's input surface, process the input, and send output to the brain's output surface. But even in the actual world, there is no necessity—not even a natural or nomic necessity—for the interstitial properties that sustain the I/O dispositions to take any particular form. A robotic brain that runs a next-generation version of ChatGPT might have the same I/O dispositions as your brain, even if the interstitial properties of the robotic brain that sustain those dispositions are very different from the interstitial properties of your brain.

[33] The "could" here designates metaphysical possibility; here I am passing over various technical wrinkles in the definition of intrinsicality, discussed in Langton and Lewis (1998).

When you have a belief, desire, or other cognitive state, your brain has certain perimetric properties; typically, these properties sustain corresponding mindlike bodily dispositions. According to perimetric functionalism, for you to have a given belief or desire is just for your brain to have the corresponding perimetric properties, and likewise for other cognitive states.

I've characterized the brain's input surface as the part of it that's directly affected by the world outside the brain (normally, the sensory anatomy of the body to which the brain is attached), and the output surface as the part that directly affects the outside world (normally, the motor anatomy of an attached body). In our brains, the input surface (thus characterized) also gets affected by activity within the brain: the connection between the sensory cortex and the rest of the brain is not a one-way street. Consequently, the total state of an actual brain's input surface at a given moment is a product of contributions from two sources: signals from outside the brain, and signals from inside the brain.

What I've been calling "the state of the input surface" at a given time is the component of the input surface's total state that is due to influences from outside the brain at that time; call this the *afferent state of the input surface*. This component of the total state and the component due to influences from inside the brain do not exist as spatially or temporally separate states of our brains. They are, so to speak, blended together in the overall state of the surface, like the shape of some dough that you press with your thumb, which is the product of the force exerted by your thumb and the resistance exerted by the dough.

Nevertheless, the components are distinguishable, like the components of an object's overall velocity contributed by different forces acting on the object. We can think of the component that is due to a tranche S of signals from sensory surfaces as a function that maps total input surface states to pairings of S with various possible intra-brain influences; this function corresponds to the afferent state of the input surface at a given moment. Likewise, we can think of the component that is due to a tranche B of signals from within the brain as a function that maps total input surface states to pairings of B with various possible extra-brain influences. If all we know is the total state of the input surface at a given moment, we can't infer which component of it is the afferent state of the surface, and which is the component due to influences from within the brain; but given the total state together with the intra-brain influences on the input surface, we can recover the surface's afferent state.

In an actual human brain, the afferent state of the input surface blends together with contributions from within the brain itself to yield the input surface's total state at a given time. In Blockhead's brain, afferent states of the input surface exist in an unblended form: all influences from within the brain come to bear downstream of the input surface. If we wanted, we could make Blockhead's brain more like a human brain, by having the Blockhead algorithm influence the states of the input nodes at each moment, based on the state of the active spreadsheet at the previous moment. This would require adding one column to each spreadsheet (containing instructions on how to modify the input nodes), and adjusting the alphanumeric contents of the other columns to ensure that the new system continued to give the same motor output as a human brain for any given regime of sensory input, but otherwise the set-up would be the same as in the original Blockhead case.

In this version of Blockhead, total states of the input surface of Blockhead's brain are a spatiotemporally unified blend of influences from inside and outside the brain, just as in an actual human brain. They are, as it were, the vector sum of influences due to signals that the input nodes receive from the body's sensory surfaces at a given moment and signals that the nodes receive from the system spreadsheet that was active at the previous moment. Presumably mainstream functionalists would not consider this version of Blockhead to be any more intelligent than the original "unblended" version.

Perimetric functionalism avoids the objections raised earlier to neo-behaviorist, black box, and Lewisian functionalism. According to perimetric functionalism, a paralytic or lock-in whose brain has the same intrinsic properties as yours is cognitively indistinguishable from you, since his brain has the same perimetric properties as yours. Likewise, an envatted brain that is intrinsically indistinguishable from your brain has the same cognitive properties as your brain, according to perimetric functionalism, since it is perimetrically indistinguishable from your brain. All of this is true, regardless of whether the paralytic, lock-in, or envatted brain belongs to a larger population: given that you believe that p, and that what sustains your corresponding belief-like dispositions are perimetric features of your brain ϕ, any subject with a brain that has ϕ also believes that p, regardless of what brain states normally sustain the relevant belief-like dispositions in members of that subject's population. According to perimetric functionalism, Joe has the same cognitive properties after the reconfiguration of his brain-body connections as he had before the reconfiguration, since his brain has the same intrinsic properties throughout, and therefore the same perimetric properties; that these properties give Joe's brain the power to sustain many mindlike dispositions that he does not in fact have is irrelevant.

Your brain is the part of you that controls your body, causing various behaviors and sustaining various mindlike bodily dispositions. Likewise, my brain is the part of me that controls my body, Jackie Chan's brain is the part of him that controls his body, and Cleopatra's brain was the part of her that controlled her body. Each of these brains has (or, had) cognitive properties, which, according to perimetric functionalism, are (or, were) perimetric properties of those brains. But not all brains sustain bodily dispositions; for example, envatted brains and the brains of total paralytics do not. Call these "unconnected brains."

If an unconnected brain has perimetric properties identical to those that are cognitive properties of some connected brain (e.g., yours), then perimetric functionalism says that the unconnected brain has those cognitive properties too; this is why perimetric functionalists can attribute cognitive properties to BIVs and paralytics where behaviorists and neo-behaviorists cannot. If an unconnected brain has perimetric properties that bear a suitable resemblance to perimetric properties that are cognitive properties of some connected brain, then perimetric functionalism says that the unconnected brain has cognitive properties correspondingly similar to those; this

is so, even if no actual connected brain has the relevant I/O properties of the unconnected brain.[34]

What if an unconnected brain has cognitive properties that bear no resemblance to any cognitive property of any connected brain known to us? According to perimetric functionalism, such a property is a perimetric property that we know nothing about, beyond perhaps that it exists. Suppose we discover an abandoned spacecraft loaded with live envatted alien brains, but no embodied aliens, and no suggestion as to what kind of bodies the brains might have controlled or how they might have controlled them. Suppose the brains are radically unlike our brains or any brains that we have encountered before, both in terms of their internal wiring and their perimetric properties. The spacecraft contains enough clues for us to deduce that the envatted entities are indeed the brains of highly evolved beings belonging to the species that built the spacecraft, so we have some basis for thinking that the things in the vats have, or are capable of having, cognitive properties. But with nothing more to go on, we can't say anything definite about what sort of cognitive properties those might be.

The moral of the story is that while perimetric functionalism tells us that any entity bearing a suitable perimetric resemblance to a brain with cognitive properties has cognitive properties that resemble that brain's, it does not tell us which entities have cognitive properties to begin with. To begin with, we have no choice but to start with our own case and work out from there as best we can. In this regard, perimetric functionalism is no different from other versions of non-analytic functionalism, which are equally powerless to say what kinds of cognitive properties the envatted alien brains might have.[35]

One virtue of behaviorism and analytic functionalism is that they *do* say which entities have cognitive properties to begin with: namely, entities with suitable mindlike dispositions, or with brains that stand in a suitable causal relation to such dispositions. For example, David Lewis can say that the reason why a certain causal state of a certain agent is a belief state is that it normally sustains belieflike bodily dispositions in members of the agent's population. In contrast, proponents of I/O functionalism have no answer, or at least no obvious answer, to the question why a certain I/O disposition of a certain agent's brain is a belief state of that agent, since, unlike Lewis and other analytic functionalists (and behaviorists), I/O functionalists don't assert an *a priori* entailment of cognitive facts by facts about mindlike behavior or its underpinnings. This is the price we pay to avoid the implausibilities of behaviorism and analytic functionalism.[36]

[34] I won't attempt to say what constitutes a suitable resemblance, but one way to establish a suitable resemblance between a perimetric property ϕ and a perimetric property ψ might be to observe that ϕ-states differ from ψ-states less than ϕ-states differ from χ-states, where χ is a perimetric property identical to a cognitive property that some connected brain has.

[35] This limitation is illusory if, as Hofweber (2023) argues, there can't be cognitive properties radically unlike ours.

[36] Wolfgang Schwarz discusses this trade-off in (Schwarz, 2015, pp. 507–12).

6 Objections to perimetric functionalism

In this section, I address two objections to perimetric functionalism. The first is analogous to David Lewis's "Mad Pain" objection to black box functionalism discussed earlier. The second is analogous to John Searle's "Chinese Room" argument against I/O functionalism of any kind.

6.1 Deep brain double-switching

Suppose we identify a part X of your brain's internal wiring that underlies the perimetric properties of your brain that sustain your belieflike dispositions regarding some proposition, p. Suppose furthermore that we have found that there are two mutually exclusive states, S_p and $S_{\sim p}$, such that X is in S_p when and only when you believe that p, and in state $S_{\sim p}$ when and only when you believe that not-p. Finally, suppose that X is currently in $S_{\sim p}$.

Now we do two things simultaneously: (1) we alter X so that X is now in S_p, and, (2) we alter the connections between X and the rest of your brain so that alteration (1) has no effect on the rest of your brain, and therefore causes no change in your brain's perimetric properties. Call this a case of "deep brain double-switching."

Perimetric functionalism implies that throughout this scenario, you believe that not-p (and do not believe that p). But, you might object, isn't it obvious that you change from believing that not-p to believing that p?

No. Imagine a magnet attached to a wooden dowel. When we spin the magnet by twisting the dowel, an electrical field results. There is a systematic correlation between the rotation of the dowel and the electrical field: no rotation, no field; faster rotation, stronger field. But it would be a mistake to conclude that we'd still get an electrical field if we twisted the dowel after loosening its connection to the magnet so that twisting the dowel no longer caused the magnet to spin. For all we know a priori, rotating the dowel *might* be sufficient for generating an electrical field, but the correlation between dowel rotation and electrical field generation in all the cases we've observed so far (in which the magnet is firmly attached to the dowel) does not give us a good reason to think that dowel rotation is sufficient for the generation of an electrical field, or to think that magnet spinning is *not* sufficient for this.

Likewise, the correlation between X being in S_p and your believing that p in all the cases we've observed so far (in which X is integrated with the rest of your brain in the usual, natural way) does not give us a good reason to think that X being in S_p is sufficient for your believing that p, or to think that your brain's operational surface being in the state it's normally in when you believe that p is *not* sufficient for your believing that p. For, in all cases observed prior to the double-switch intervention, X was observed to be in S_p when and only when your brain had the perimetric properties that sustained the believing-that-p-like dispositions that were our basis for think-

ing you believed that p. (States of X are analogous to states of the dowel; perimetric properties are analogous to states of motion of the magnet.)[37]

So it's an open question, thus far, whether the double-switch would induce any mental change in you. How might we settle it?

Not by soliciting testimony from you. Whatever testimony you provide, it's the same regardless of whether your beliefs change; this follows from the fact that the double-switch has no effect on your brain's perimetric properties, and therefore no effect on your verbal behavior or dispositions.

It remains to consider whether *you* could have credible evidence that the double-switch induced a cognitive change: evidence that you cannot communicate to anyone else.

If you could have such evidence, it would presumably come from introspection. Introspection can, of course, provide you with evidence about your cognitive states, but it's important to realize that this is defeasible evidence. (This is true, even if introspection can provide us with indefeasible evidence of our occurrent *phenomenal* states.) We know that introspection provides only defeasible evidence of our cognitive states, because it often happens that people's introspective judgements about their own cognitive states are false. This kind of error is well-documented in the psychology literature. There are racists and misogynists who falsely think they believe in race and gender equality, spouses of unfaithful partners who falsely think they believe their partners are faithful, and philosophers who falsely think they doubt there is an external world.[38]

Since introspection of cognitive states is fallible, even if you did have an (incommunicable) introspective belief that you believed that not-p in the double-switch scenario, that wouldn't give you a compelling reason to think that you really believed that not-p. So far, we haven't seen any reason to think you would have such an introspective belief in a double-switch scenario. But suppose you would. Then you would have a false introspective belief about your own beliefs, according to perimetric functionalism. What might explain this?

Here is a possible explanation. When you use introspection to find out whether you believe that not-p, your brain scans its simplest and most readily accessible part whose states normally covary with your p-related beliefs. In the case we're considering, this part is X (or perhaps some proper part of X). It makes sense for introspection to work this way, assuming that it costs the brain less to detect whether X is in state S_p than to detect whether the brain's operational surface is in the perimetric state that normally sustains believing-that-p-like bodily dispositions in you. And this is a reasonable assumption, since the only way the brain can detect whether it's in the relevant perimetric state is by gathering information about the parts of the brain's

[37] For analogous reasons, the fact (if it is a fact: cephalopods suggest otherwise) that the brains of humans and other intelligent animals have similar internal wiring doesn't show that intelligence is a feature that animals have in virtue of having brains with a certain kind of internal wiring. Since the brains of intelligent animals also have similar I/O properties, similarities among intelligent animals can't decide between mainstream and I/O functionalism.

[38] For relevant empirical data, see Nisbett and Wilson (1977), Nisbett and Ross (1980), and the extensive literature corroborating these.

internal wiring that normally mediate the dependence of various motor neurons on various sensory neurons, and those internal parts include X.

Ordinarily, this relatively low-cost way getting information about your cognitive states works reasonably well, since ordinarily the states of X covary with the perimetric states that are (according to perimetric functionalism) your p-related belief states. But when X is prevented from influencing your brain's operational surface in the normal way, as in the double-switch scenario, X's states no longer covary with your p-related belief states (since they no longer covary with your brain's relevant perimetric properties), and introspection no longer provides reliable evidence about your beliefs regarding p.

The temperature gauge on my car's dashboard is usually a reliable source of information about the thermal state of my car's engine. But what the gauge *directly* reflects is the state of a thermometer attached to the engine's block. If the mercury in the thermometer falls, the dashboard gauge will indicate a decrease of engine temperature, even if the mercury falls because the thermometer has become decoupled from the engine so that its states no longer reflect the thermal states of the engine. Something analogous happens if, during the double-switch, introspection reports that you believe that p, based on the fact that X is in state S_p: normally, X being in S_p is a reliable indication that you believe that p, but when X is decoupled from your brain's operational surface (as in the double-switch scenario), this is no longer the case.

In summary: (1) there is no obvious reason to think that deep brain double-switching would induce cognitive change, (2) it's impossible for there to be third-party evidence that deep brain double-switching induces cognitive change, and, (3) any introspective belief that double-switch-induced cognitive change had occurred would be (a) defeasible, and, (b) discredited by a plausible account of how introspection of cognitive states works.[39]

6.2 Nesting

Let Your Brain = your actual brain with its actual perimetric properties. In reality, your body's sense organs send signals directly to Your Brain's input surface, and Your Brain's output surface sends signals directly to your body's motor anatomy. But imagine an alternative reality in which no body sends signals directly to Your Brain's input surface, and Your Brain's output surface doesn't send signals directly to any body. Instead, the sense organs of a body indistinguishable from Jackie Chan's send signals directly to a structure S_i, which then sends signals directly to Your Brain's input surface, and Your Brain's output surface sends signals directly to a structure S_o, which then sends signals directly to the Chan body's motor anatomy. Finally, imagine that the perimetric properties of the entity that comprises S_i and S_o are identical to the perimetric properties of the operational surface of Jackie Chan's brain. (As

[39] Considerations like these might have led Wittgenstein to his behaviorist or quasi-behaviorist theory of mind: see (Wittgenstein, 1958/2001, §258, §293, §580, and p. 177).

already stipulated, the perimetric properties of Your Brain's operational surface are identical to your brain's actual perimetric properties.)[40]

Let's give the name "Nestor" to the being that results from the described interpolation of S_i between the Chan body's sense organs and Your Brain's input surface and S_o between Your Brain's output surface and the Chan body's motor anatomy.

At first, Nestor might seem to pose a problem for perimetric functionalism, similar to the problem that agents like Joe pose for neo-behaviorist functionalism. Recall that Joe was a person whose brain-body connections were reconfigured so as to result in his body's having very different mindlike dispositions from what it had originally, despite Joe's brain undergoing no intrinsic change (and therefore no change in I/O properties). Neo-behaviorist functionalism forces its proponents to say, implausibly, that reconfiguring Joe's brain-body connections results in Joe's undergoing major cognitive changes, despite the lack of any intrinsic changes in his brain. Don't perimetric functionalists similarly have to say that making the changes required to create Nestor from your existing organism would result in your undergoing major cognitive changes, such as acquiring a knowledge of Chinese, despite the lack of any changes in your brain's intrinsic properties?

No. Nestor is what we might call a nested cognitive agent. He is, in fact, two agents. One is the agent whose brain's operational surface has the same perimetric properties as your brain actually has; call this Agent 1. The other is an agent cognitively indistinguishable from Jackie Chan: this is an agent whose brain's operational surface comprises S_i and S_o and has perimetric properties identical to those of Jackie Chan's brain; call this Agent 2. The case is peculiar, for two reasons: first, because Agent 1's brain serves as the internal wiring of Agent 2's brain, and, second, because only one of the agent's brains—that of Agent 2—directly controls a human body. Because of this, Agent 1 lives in a state of constant deception; for example, Agent 1 believes that he or she has a body with eyes that are now scanning a philosophy article, when this is not in fact the case. (Maybe the Chan body's eyes are currently scanning a movie set in Hong Kong; whatever it's doing, it's not *your* body.) However, this is compatible with perimetric functionalism, which entails only that Agent 1 has the same beliefs, desires, etc. as the actual you, but not that those beliefs have the same truth-values as your actual beliefs.

7 Perimetric functionalism about consciousness

So far, I've argued that our cognitive properties are perimetric properties of our brains. Like others who are skeptical about materialist theories of consciousness, I do not think that our *phenomenal* properties are physical properties of our brains, perimetric or otherwise. However, I do think that perimetric properties *determine* phenomenal properties, in the sense that it is naturally or nomically (though not metaphysically) necessary that any brain with the same perimetric properties as the brain of a given conscious being has conscious experience indistinguishable from that being's.

[40] This scenario is reminiscent of (and inspired by) the setup that John Searle describes in the second chapter of Searle (1984).

Though the argument that follows does support the claim that phenomenal properties nomically or naturally supervene on perimetric properties, its immediate aim is to establish that if a brain has the same perimetric properties as yours, there's as much reason to think it has the same phenomenal properties as your brain as there is to think that a brain physically indistinguishable from yours has the same phenomenal properties as your brain. This epistemic proposition is already enough to guide us in our deliberations about whether to judge that various artificial systems are sentient.

That said, if we can establish the epistemic proposition, that provides the material for an abductive argument for the nomological proposition: (1) evidence of perimetric equivalence is on a par with evidence of physical equivalence, when it comes to attributions of phenomenal properties; (2) the best explanation of this is that phenomenal properties naturally supervene on perimetric properties; therefore, (3) phenomenal properties naturally supervene on perimetric properties. For brevity's sake, I present the argument that follows as an argument for (3).

The argument is an adaptation of David Chalmers's dancing qualia argument. Chalmers uses this kind of argument to show that our phenomenal properties naturally supervene on our brains' global functional organizations; I use it to show that our phenomenal properties naturally supervene on our brain's perimetric properties, which are a proper subset of the properties that characterize our brains' global functional organizations.[41]

Imagine a partial Blockhead—Blocky, to give him a name—whose brain is physically indistinguishable from yours, except in one respect: where your brain has a natural, organic visual cortex, Blocky's brain has an automated Excel workbook. The workbook connects to the rest of Blocky's brain the same way your visual cortex connects to the rest of your brain, and the workbook has been composed in such a way that its dispositions to send output to the rest of Blocky's brain given input from outside the workbook (e.g., from Blocky's eyes) are the same as your natural cortex's dispositions to send output to the rest of your brain given input from outside the cortex (e.g., from your eyes).

Blocky is perimetrically equivalent to you: his brain has exactly the same intrinsic I/O properties as yours. The difference between your brains is purely one of internal wiring. If my arguments for identifying your cognitive properties with perimetric properties of your brain are sound, it follows that Blocky is cognitively equivalent to you. In particular, he has the same beliefs as you.

Now assume, for reductio, that it is *not* naturally necessary that Blocky has visual experience phenomenally indistinguishable from yours—e.g., that it's not naturally necessary that Blocky has visual experience with the same phenomenal color-scheme as yours, rather than an inverted color-scheme, or not naturally necessary that Blocky has any visual experience at all. We show that this assumption leads to absurdity, as follows.

Suppose we arrange things so that at the flip of a switch, two things happen simultaneously: Blocky's workbook starts to receive input from and send output to the extra-visual part of your brain (that is, the part of your brain that excludes your natural visual cortex) rather than the extra-visual part of Blocky's brain, and your natural

[41] For Chalmers's argument, see (Chalmers, 1996, pp. 247–75).

visual cortex starts to receive input from and send output to the extra-visual part of Blocky's brain rather than the extra-visual part of your brain. In effect, flipping the switch causes Blocky's workbook to replace your natural cortex and vice versa.

If it's naturally possible for Blocky not to have conscious visual experience, or to have conscious visual experience color-inverted relative to yours, then it's naturally possible that flipping the switch causes Blocky to start having conscious visual experience, or causes the phenomenal color of his visual experience to change from its original scheme to a color-inverted scheme; in that case, repeatedly flipping the switch back and forth would either cause Blocky to repeatedly lose and regain visual consciousness, or cause his color qualia to "dance" back and forth from one color scheme to another.

However, flipping the switch has no effect on the perimetric properties of Blocky's brain, since the workbook has the same input-output properties as the natural cortex, in relation to the rest of the brain. Given the identification of our cognitive properties with our brain's perimetric properties, it follows that flipping the switch has no cognitive effect. Before we started flipping the switch, Blocky did not believe that he was repeatedly losing and regaining visual consciousness, or that his visual qualia were dancing, and (we may suppose) he did believe that he was having uninterrupted and chromatically stable visual experience. Since flipping the switch has no cognitive ramifications, when we start flipping it, Blocky continues not to believe that he repeatedly loses and regains visual consciousness or that his visual qualia are dancing, and continues to believe that he has uninterrupted and chromatically stable experience.

But if Blocky doesn't believe that his visual qualia are changing or intermittently disappearing, and does believe that they are not changing and not intermittently disappearing, that is a very powerful reason to conclude that his visual qualia are not changing or intermittently disappearing. Introspection of phenomenal states may be fallible, but it's not so fallible that a sane, sober subject can fail to notice sudden dramatic changes in the phenomenal quality or quantity of his experience, or believe that his experience remains the same in phenomenal quantity and quality when in fact it changes dramatically in those respects. It does not seem too strong to say that it is naturally necessary that sane, sober subjects are not mistaken when they believe that their occurrent conscious experience does not undergo sudden phenomenal changes like this.[42]

So after we first flip the switch, Blocky's visual phenomenology is the same as before we flipped it. But when we first flip the switch, Blocky's brain becomes physically indistinguishable from your brain prior to the switch-flipping: his brain now consists of a natural visual cortex integrated in the natural way with the rest of a natural brain. Assuming (plausibly) that it is naturally necessary that physically indistinguishable brains have phenomenally indistinguishable experience, it follows that it is naturally necessary that Blocky's pre-switching visual experience was phenomenally indistinguishable from yours. This completes the reductio of our initial assumption that it is *not* naturally necessary that Blocky has visual experience phenomenally indistinguishable from yours. Here is the whole argument:

[42] Consider: are you less certain that your qualia aren't now dancing than that $E = mc^2$?

1. It's naturally possible that Blocky's visual qualia change when we flip the switch. (Assume for Reductio)
2. It's not naturally possible that Blocky's beliefs change when we flip the switch.
3. If it's not naturally possible that Blocky's beliefs change when we flip the switch, but it is naturally possible that Blocky's visual qualia change when we flip the switch, then it's naturally possible that Blocky mistakenly believes that his visual qualia go unchanged.
4. So it's naturally possible that Blocky mistakenly believes that his visual qualia go unchanged. (1, 2, 3)
5. But it's not naturally possible that Blocky mistakenly believes that his visual qualia go unchanged. (Contradiction: 4, 5)

6. It's not naturally possible that Blocky's visual qualia change when we flip the switch. (1–5, by Negation Introduction)
7. If it's not naturally possible that Blocky's visual qualia change when we flip the switch, then it's naturally impossible for a being's visual qualia to differ from those of someone whose brain has the same perimetric properties as that being (i.e., a "perimetric isomorph" of that being).
8. So, it's naturally impossible for perimetric isomorphs to have different visual qualia. (6, 7)

Similar arguments apply to other, non-visual phenomenal properties; thus we may conclude that it is naturally necessary that brains with the same perimetric properties as our brains have the same phenomenal properties as our brains. Since Blockhead (original all-workbook version) has a brain with the same perimetric properties as an ordinary human brain, it follows that it's naturally necessary that if Blockhead existed, he would have conscious experience indistinguishable from that of an ordinary human being.

Chalmers's version of the argument says "global functional properties" where my version says "perimetric properties," and "functional isomorphs" where mine says "perimetric isomorphs." This reflects the fact that in Chalmers's thought-experiment, a natural visual cortex is paired with a silicon device that duplicates both the I/O architecture and the internal causal structure of the natural cortex, where in my thought-experiment, the pairing is with a device that duplicates only the natural cortex's I/O architecture.

In both versions, the controversial premises are (2) and (5).

Regarding (5), Chalmers acknowledges that it might be possible to believe mistakenly that one's qualia go unchanged, even if one's qualia change in a very subtle way. However, as he points out, this does little to diminish the force of the argument; at most, we'd have to modify it to conclude that it's naturally impossible for functional isomorphs to differ in their visual qualia except maybe in very subtle ways (and likewise for other types of qualia, including, e.g., pain qualia).[43] In my version of the argument, the modified conclusion would be that it's naturally impossible for perimetric isomorphs to differ in their visual qualia except maybe in very subtle ways (and likewise for other types of qualia).

[43] See (Chalmers, 1996, pp. 271–72).

The main objection to (2) as it occurs in Chalmers's argument is that Chalmers doesn't provide much support for it, merely remarking that there is "simply no room" in functionally isomorphic brains for any difference of beliefs between them.[44] In contrast, I have provided arguments for the claim that perimetrically isomorphic brains have the same beliefs: the causal and evolutionary arguments, supported by subsequent arguments for the cognitive irrelevance of properties of the brain distinct from its I/O properties (such as unscripted adaptiveness and energy efficiency). Since functional isomorphs are also perimetric isomorphs, these arguments also support Chalmers's version of (2). But my version of the dancing qualia argument is more consequential, since it implies that it's much easier to build a conscious machine than mainstream functionalists like Chalmers suppose: it requires only building a machine with a suitable degree of perimetric resemblance to human brains, rather than a machine that also resembles human brains in its internal wiring.

8 Conclusion

The current Zeitgeist in the philosophy of mind is markedly conservative. In an extended white paper on machine consciousness released in August 2023, the co-authors—nineteen respected philosophers, psychologists, neuroscientists, and computer scientists—take conservatism as a methodological axiom, dismissing viewpoints left of mainstream functionalism almost out of hand.[45] Most experts are more liberal than Ned Block, but nearly all are more conservative than I/O functionalists.[46]

This paper has challenged the prevailing orthodoxy. The time may be closer than people realize when the machines rolling off the high-tech assembly line are genuinely intelligent, sentient agents. We would do well to prepare ourselves for this by reflecting now on whether the orthodox conservative requirements for sentience and intelligence are appropriate, or whether, as I have argued, they are excessively stringent. Failing to do so could have serious consequences for both us and our creations.[47]

Acknowledgements None.

[44] (Chalmers, 1996, pp. 258, 269). Someone like Ned Block is apt to reply that the inorganic character of a silicon brain makes room for a cognitive difference between that brain and an organic isomorph of it.

[45] See Butlin et al. (2023). The closest the authors come to giving an argument against more liberal views is the comment that "AI systems can be trained to mimic human behaviours while working in very different ways." (Butlin et al., 2023, p. 4)

[46] Cappelen and Dever (2025) is a notable exception, though only with respect to cognition.

[47] I shared earlier versions of this paper with audiences at Sacramento State University's Center for Practical & Professional Ethics, Virginia Commonwealth University, Virginia Tech, and the University of Vermont; I gratefully acknowledge the feedback I received on those occasions, with special thanks to Kyle Swan, Eugene Mills, Kelly Trogdon, Daniel Hoek, Louis deRosset, and Matthew Weiner. I also thank David Braddon-Mitchell, Riley Harris, Frank Jackson, Robert Kirk, and William Lycan for their valuable input. I conducted most of the research for this paper while on sabbatical at the University of North Carolina at Chapel Hill; my thanks to the UNC Philosophy Department for hosting me during that time. As always, my colleagues at NUS served as invaluable sounding-boards. Last but not least, thanks to the reviewers for this journal, whose comments prompted many improvements.

Author contributions Single-author.

Funding None.

Data availability NA.

Declarations

Competing interests None.

References

Armstrong, D. (1968). *A materialist theory of the mind*. Routledge & Kegan Paul.

Bailey, A. (2016). Composition and the cases. *Inquiry*, *59*(5), 453–470. https://doi.org/10.1080/0020174X.2015.1122550

Block, N. (1981). Psychologism and behaviorism. *The Philosophical Review*, *90*(1), 5–43. https://doi.org/10.2307/2184371

Block, N. (2026). Can only meat machines be conscious? *Trends in Cognitive Sciences*, *30*(4), 298–308. https://doi.org/10.1016/j.tics.2025.08.009

Braddon-Mitchell, D., & Jackson, F. (2007). *The philosophy of mind and cognition*. (2nd edn.). Blackwell.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv*.

Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese*, *200*(6), 1–31. https://doi.org/10.1007/s11229-022-03524-1

Cappelen, H., & Dever, J. (2025). *Going whole hog: A philosophical defense of AI cognition*. arXiv.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.

Cobham, A. (1965). The intrinsic computational difficulty of functions. In: Y. Bar-Hillel (Ed.). *Logic, methodology and philosophy of science: Proceedings of the 1964 International Congress* (pp. 24–30). North-Holland Publishing Company.

Davies, M. (1992). Aunty's own argument for the language of thought. In J. Ezquerro, & J. M. Larrazabal (Eds.). *Cognition, semantics and philosophy: Proceedings of the first international colloquium on cognitive science* (pp. 235–271). Springer.

Dennett, D. C. (1981). *Brainstorms: Philosophical essays on mind and psychology*. MIT Press.

Feuillet, L., Dufour, H., & Pelletier, J. (2007). Brain of a white-collar worker. *Lancet*, *370*(9583), 262. https://doi.org/10.1016/S0140-6736(07)61127-1

Fodor, J. (1987). *Psychosemantics*. MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Godfrey-Smith, P. (2023, October). Nervous systems, functionalism, and artificial minds. A talk given online to the NYU Mind, Ethics, and Policy Program.

Hofweber, T. (2023). *Idealism and the harmony of thought and reality*. Oxford University Press.

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

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

Langton, R., & Lewis, D. (1998). Defining 'intrinsic'. *Philosophy and Phenomenological Research*, *58*(2), 333–345. https://doi.org/10.2307/2653512

Lewis, D. (1980). Mad pain and Martian pain. In N. Block (Ed.), *Readings in philosophy of psychology* (Vol. 1, pp. 216–222). Harvard University Press.

Lycan, W. G. (1987). *Consciousness*. MIT Press.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*(3), 113–126. https://doi.org/10.1145/360018.360022

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Prentice-Hall.

Nisbett, R., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231
Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology*, *47*(1), 44–50. https://doi.org/10.1111/j.2044-8295.1956.tb00560.x
Pritchard, D. (2007). Anti-luck epistemology. *Synthese*, *158*(3), 277–297. https://doi.org/10.1007/s11229-006-9039-7
Putnam, H. (1980). The nature of mental states. In N. Block (Ed.), *Readings in philosophy of psychology*, (Vol. 1, pp. 223–31). Harvard University Press.
Ryle, G. (1949). *The concept of mind*. University of Chicago Press.
Schwarz, W. (2015). Analytic functionalism. In B. Loewer, & J. Schaffer (Eds.), *A companion to David Lewis.* (pp. 504–518). Wiley Blackwell.
Schwitzgebel, E. (2025). AI and consciousness. arXiv.
Schwitzgebel, E., & Pober, J. (2025). The Copernican argument for alien consciousness; the mimicry argument against robot consciousness. arXiv.
Searle, J. R. (1984). *Minds brains and science*. Harvard University Press.
Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review*, *68*(2), 141–156. https://doi.org/10.2307/2182164
Sosa, E. (1999). How to defeat opposition to Moore. *Noûs*, *13*, 141–153. https://doi.org/10.1111/0029-4624.33.s13.7
Tiku, N. (2022, June 11). The Google engineer who thinks the company's AI has come to life. *The Washington Post*.
Williams, J. R. G. (2019). *The metaphysics of representation*. Oxford University Press.
Williamson, T. (2000). *Knowledge and its limits*. Oxford University Press.
Wittgenstein, L. (1958/2001). *Philosophical investigations: The German text, with a revised English translation*. Blackwell Publishing.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.