



Multiple realizability and the spirit of functionalism

Rosa Cao¹ 

Received: 14 April 2021 / Accepted: 25 November 2021 / Published online: 1 December 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

Multiple realizability says that the same kind of mental states may be manifested by systems with very different physical constitutions. Putnam (1967) supposed it to be “overwhelmingly probable” that there exist psychological properties with different physical realizations in different creatures. But because function constrains possible physical realizers, this empirical bet is far less favorable than it might initially have seemed, especially when we take on board the richer picture of neural and brain function that neuroscience has been uncovering over the past forty years, in which all sorts of brain activities play crucial roles beyond all-or-nothing electrical signaling. Because of its evolutionary history, the brain’s metabolic and informational processes are inextricably intertwined. The resulting complex integrated functions impose more constraints on possible realizers than the clean, single-purpose functions usually cited as examples in discussions of multiple realizability—with ramifications for the functionalist and computationalist foundations of cognitive science, artificial intelligence, and the philosophy of mind.

Keywords Embodiment · Neurons · Artificial intelligence · AI · Mind · Reductionism · Reduction · Emergence · Representation · Functionalism · Brain · Identity theory · Behaviorism · Biology · Neuroscience · Metabolism · Constraints · Explanation · Consciousness

1 Introduction

Multiple realizability is the thesis that the same kinds of mental capacities can be manifested by systems with very different physical constitutions. Putnam (1967) supposed it to be “overwhelmingly probable” that there exist psychological proper-

✉ Rosa Cao
rosacao@stanford.edu

¹ Dept of Philosophy, Stanford University, Stanford, CA, USA

ties with different physical realizations in different creatures. Fodor (1974) thought multiple realizability “increasingly likely” in automata, and took it as support for the autonomy of psychology. It remains a common claim that a functional duplicate of a person’s mind could be realized in silicon, or some other non-biological substrate. The appeal of functionalism in philosophy and cognitive science depends in significant part on the plausibility of multiple realizability as a favorable empirical bet about the nomological or physical possibility of alternate realizations of mental states or properties.

I want to show that this bet is far less favorable than it might initially have seemed. Far from freeing us from worrying about substrates, functional roles *constrain* physical realizations. It is only in virtue of some of its physical properties that any particular component is able to play its particular functional role. Conversely, having to play a particular functional role imposes constraints on what kinds of material stuff can do the job. Ned Block (1997) calls this the “Disney principle”—but it’s really more of an anti-Disney principle: you can’t have dancing cups or plates made out of the stuff that normal plates and cups are made of.

What is the relevant sense of “can’t” here? Discussions of multiple realizability have often been concerned primarily with *metaphysical* possibility.¹ But if we care about functionalism as a guide to cognitive science and artificial intelligence research, big science connectomics or even the fringe hope of uploading one’s mind into a computer, then *nomological* possibility (in the relatively conservative sense of physically-possible-in-this-world-if-we-had-enough-resources) is the more relevant standard.

This paper can be seen as an cautionary note, joining a chorus of work over the last few decades, all aimed at downgrading our confidence in intuitions about how the mind works that abstract away from the messy details of physical implementation in a biological organism.² Prominent earlier arguments against this canonical version of MR include Bechtel & Mundale (1999), Polger (2004, 2008), and Shapiro (2000, 2004, 2008), describing problems with traditional formulations (and putative instances) of multiple realization.

Bechtel and Mundale argue that many of the standard examples cited as cases of multiple realizability don’t really qualify—and that the illusion of multiple realizability comes from comparing a fine-grained characterization (e.g. that of a detailed neurophysiological state) with a coarse-grained one (e.g. “pain”).³ Polger and Shapiro,

¹ By contrast, if what is at issue is only the metaphysical question of how to analyze our mental terms, then perhaps nomological constraints can be left aside.

² As such, it is part of a tradition starting from the pioneering neurophilosophy of the Churchlands to the naturalistic philosophy of mind and neuroscience of Bechtel, Bickle, Craver, Klein, Roskies, Sullivan and many others, especially the influential work of Polger and Shapiro—who, though clearly not all in agreement, remind us of the many reasons to think that the body and its brain do not just house the mind like a fleshly vat, but rather, constitute it, and that the many gory details are not just irrelevant biology, but profoundly relevant to the functioning of the system. Bickle (2010) provides a survey and critique of the earlier contributions to this debate.

³ Bechtel and Mundale contend that if we considered brain architectures at as coarse-grained a level as we casually speak of mental states like “pain”, we would find that the architectures had as much in common as the mental states, and therefore that they would constitute tokens of the same realization. At a finer grain of description, there are many differences between human versions of such components and those

separately and together, have argued that multiple realizability must be understood as a contrastive question: always relative to the relevant taxonomies at the level of functional description and the level at which realizers are characterized. And they, too, stress the importance of paying attention to how functions and constraints interact in assessing the plausibility of MR.⁴

Given these and the many other arguments in this vein, it is surprising that the dominant intuition that I target here is one that has received relatively little scrutiny within philosophy. This target is the neural doctrine: that the only functionally relevant aspect of the brain, what gives it its behavioral capacities and gives us our mental states, is neuronal activity, and in particular, patterns of action potentials. All sensory inputs are transduced by the peripheral sense organs into electrical signals, which propagate through the brain, which is made of neurons. At the other end, electric signals are sent to the muscles, which translate them into motor outputs. If this *neuron doctrine* is true, then the relevant functional organization to be replicated for an alternative realization of the mind is just a complicated and dynamic system of electric impulses, which then set off other signals in appropriate ways—a picture that fits satisfyingly into the traditional functionalist schema.

But there are reasons to be suspicious of so conveniently clean a picture as that painted by the neuron doctrine.⁵ If we take on board the richer picture of neural and brain function that neuroscience has been uncovering over the past forty years, in which all sorts of brain activities beyond electrical signaling play crucial roles in the functional architecture and operation of the system, multiple realizability will seem less obvious, at least at the level of neural function and organization.

Here's the basic argument: Functions impose constraints on their realizers. Complex integrated functions impose more constraints than ones that can be decomposed into simpler independent functions. Mental functions are both complex and thickly integrated with physiological functions. So the functional organization of the brain at the cellular level exhibits characteristics that make it unlikely to be multiply realizable at that level of description.

There are many distinct philosophically interesting multiple realizability claims—at the personal level, at the level of subpersonal processing, and at the fine-grained neural level. I will start by arguing against the nomological possibility of MR on the neural level, both because it is a familiar case, and because it undergirds the presumption of MR more generally. MR at the neural level entails MR at higher levels, under the plausible assumption that psychological properties supervene on

of the octopus—but Bechtel and Mundale's claim is that at a finer grain of description there will also be differences between the functional profile that characterizes my pain and the one that characterizes that of the octopus.

⁴ See especially chapters 3 and 4 of Polger and Shapiro (2016) for careful developments of a number of related considerations that push them towards updated versions of a type-identity theory. They maintain that the burden of proof should be on defenders of multiple realizability, not its detractors. See Aizawa & Gillette (2009), and also Fang (2018) for counterpoint. There are also alternative characterizations of MR that allow for realizations that exhibit differences in functional organization to nevertheless count as alternative realizations of the *same* functional profile (e.g. Piccinini & Maley 2014). My argument is directed only at the canonical formulation from Putnam and Fodor inherited by mainstream philosophy of mind and much of cognitive science.

⁵ See Bullock (2005) for a survey of these reasons, and Cao (2014) for a philosophical perspective.

neural properties. Add to this the presumption that neurons are relatively simple entities—transistor-like all-or-nothing switches, for instance—and it becomes clear why people have taken it to be *obvious* that multiply realization is true at all levels of functional organization.

Undermining the presumption that neurons are easily replaceable by non-biological substitutes will cast doubt on the claims of MR proponents more generally. I'll then give some reasons for why I think those considerations against low-level MR may generalize to more macro levels of organization, before examining the implications for non-chauvinist functionalism more generally.⁶

2 Why neurons—and brains—are hard to replace

2.1 The neural replacement scenario

In one vivid example, Chalmers (1995) describes a scenario in which a person's neurons are replaced, one by one, with silicon functional duplicates, until all the work of the brain is performed by an entirely artificial system.⁷ The resulting alternative realization is supposed to be an *exact* functional duplicate of the original over a short period of time.

Its replacement is a silicon chip that performs precisely the same local function as the neuron. We can imagine that it is equipped with tiny transducers that take in electrical signals and chemical ions and transforms these into a digital signal upon which the chip computes, with the result converted into the appropriate electrical and chemical outputs. As long as the chip has the right input/output function, the replacement will make no difference to the functional organization of the system.

It is true that *if* the silicon chip has the right input/output function, then the replacement will make no difference to the functional organization and behavior of the system, since that is just what the functional organization is supposed to capture. The key question is: *can* a silicon chip do these things, given the nomological constraints?

⁶ With the advent of neural network models and concomitant advances in artificial intelligence, it might seem that these considerations have already been found irrelevant. But in fact, the way in which those models succeed, when they do succeed, serves to underline the importance of the constraining relationship between functional demands and physical implementation.

⁷ Pylyshyn (1980) and others have introduced very similar thought experiments. Pylyshyn writes: "...if more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output function of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now ...". I should also note that while Chalmers was concerned to make a point about conscious experience, I'm concerned here only with whether the functional organization of the brain is multiply realizable, and not whether the organizational invariance thesis about conscious experience is true. (Everything in this section will be compatible with a thesis of organizational invariance, since invariance claims only that same organization entails same mental state, with no commitments about whether that functional organization can be realized by different physical systems).

To begin with, notice that by focusing on *neuron* replacement, the replacement theorist is tacitly accepting whatever turns out to be the best scientific account of the brain's functional units — not just picking them out, but also characterizing their functional role. That means that as our best scientific account of neural function improves, proponents of MR should update their bets accordingly. I would add that since these are empirical claims, the burden of proof should lie at least as heavily on proponents of MR as on its opponents.

In order to specify the input-output function of a silicon chip replacement, we need to specify the input-output function of the neuron that it is replacing. On our best neuroscientific understanding, the input-output function of a neuron is *not* exhausted by electrical excitation on the input side and an all-or-nothing signal on the output side. At the very least, the I/O profile of a successful replacement neuron should be sufficient to account for the original's contribution to behavior and to the electrical excitation of other neurons—and what is required for *that* is considerable more complex: spatially varying, history-dependent sensitivity to a variety of biochemical impingements on the input side, and then outputs that include self-modification⁸ and changes in connectivity as well as action potential generation and neurotransmitter release.

These requirements provoke constraints that are *not* merely practical or technological ones. Rather, they are nomological constraints of the same kind as those that tell us that that water is not superconducting at room temperature, or that the biochemical constitution of DNA molecules limits their possible three dimensional configurations. I'll examine the requirements and constraints one at a time below.

2.2 Why neural replacement is hard

a) *Spatiotemporal characteristics of a neuron's spiking responses.*

Neurons are not simple bipolar devices. Rather, they are spatially extended, with broad dendritic arbors (these may span a volume hundreds of microns across, while the cell body is only about 40 microns across) and often significant axon lengths (e.g. across the corpus callosum). As a result, inputs to the cell are spatially distributed in functionally significant ways; how they combine to produce outputs (which are also spatially distributed) is related to their relative arrival times as well as intracellular propagation speeds. Furthermore, those conduction speeds change constantly, since the conductance of the cell membrane changes depending on what is currently happening.⁹

Size matters more generally because spatial locations matter for determining which interactions happen and when—changing the size and shape of the replace-

⁸ It seems reasonable to generalize the idea of the “output” here to include internal changes, because when the unit changes its own state, it also changes its future dispositions to respond to input, and thus its input-output profile considered over a longer time interval.

⁹ To complicate things further, from the point of view of action potentials, the relevant functional unit may not be the whole neuron, but rather a single dendrite, or perhaps a group of them, which propagate local spikes resulting in local neurotransmitter release, but which rarely if ever have any effect on the cell body or axons. See London and Hausser (2005) and Mel (2007) for reviews.

ment units will change whether and how they interact with the rest of the system, and thus potentially alter the functional organization of the area.

Neural responses are also history-dependent—and it is part of their function that this is the case. Only then can they make dynamic adjustments to changing circumstances. Thus a neuron will respond differently to identical inputs depending on what its recent history has been, and very small differences in history can have large impacts on the output at a given time. So both morphology (or equivalently, just the right configuration of delay and amplification or attenuation elements) and history-dependent features of the neural I/O profile would need to be engineered into the chip.

So far this is all the sort of thing that could, in principle, be calculated or simulated by other means, given enough time and resources—but it is not obviously the sort of the thing that could be calculated *quickly* enough, using a device *small* enough to replace an actual neuron.¹⁰ When we consider simulating molecular interactions at the same space and time scale at which they actually fluctuate, it might well turn out that there is no alternative that can work as quickly as implementing the actual analog process itself.¹¹ As Aaronson (2011) argues, sometimes complexity limitations are principled ones, rather than merely practical ones.¹²

We might have thought that silicon (where “silicon” is a shorthand for whatever combination of materials make up our artificial chip) provides naturally faster speeds than neural conduction. After all, we hear all the time that silicon based hardware is “faster” than biological wetware, and this is clearly true for certain operations. Since it’s much easier to introduce delays than to speed up a process already running at its maximum speed, we might also have then supposed that there is plenty of excess processing capacity in a silicon chip to model or mimic whatever idiosyncrasies occur in a biological cell. And if the cell’s input/output function were relatively simple, I’d agree with this assessment. If, however, the cell is more like a sophisticated analog computer that is fully exploiting many of its physical features in order to perform complex “computations”, then that assessment seems too optimistic.

I have until this point glossed over distinctions between different kinds of functionalism. But it is worth emphasizing that what we care about is a dispositional input-output function that includes failure profiles and system properties such as robustness to various kinds of noise or interference. These are matters on which a

¹⁰ Speed matters because behavior is partly individuated temporally, and the functional significance of many mental or psychological states relies on response speeds and the right kind of temporal coupling either to other processes or to the environment. For more on the role of speed, see Dennett’s *Fast Thinking* (1987).

¹¹ When we consider the massive computational cost of simulating a single protein-protein interaction for less than a millisecond, it should give us pause about the plausibility of large-scale simulations of molecular dynamics over physiologically relevant periods of time. For a discussion of the constraints faced by such simulations, see Hollingsworth and Dror (2018). Computation may be universal in theory, but enough orders-of-magnitude differences between the simulation efficiency and the original can pile up to make some simulations nomologically impossible in the time and space available. “Quantum supremacy” might be thought of as an extreme example of the constraints imposed on the computational medium by a computational task.

¹² “Why Philosophers Should Care About Computational Complexity”, Scott Aaronson (2011), Electronic Colloquium on Computational Complexity, Report No. 108.

crude machine functionalism would be silent, but ones that matter for a modern functionalism given in terms of a full specification of the causal roles of the components of a system. Moreover, insofar as the input-output profile of the whole system is characterized in terms of actual interactions with its environment, the size and speed of its components will be factors that are relevant to its functional organization. Of course it is *relative* size and speed that matter (relative to other processes in the system and the environment), but we are holding the environment fixed in these replacement scenarios. That is, if we are long-armed functionalists, then facts about the environment will *anchor* realizations to particular spatiotemporal scales, both externally and internally, and that in turn constrains possible realizers.¹³

b) *Transducers and chemical signaling*

We also have to take into account the many *kinds* of input: excitatory, inhibitory, and other biochemically mediated modulations that affect the particular way that excitatory and inhibitory signals are combined. Action potentials depend on tens of thousands of selective ion channels, with dynamics that are sensitive to the local biochemical milieu. In order to duplicate the functional sensitivity of a single neuron to its neighbors, we would need a large number of sensors/transducers.

How many? We would need one for every ion channel (of which there are tens of thousands in a cellular membrane, each with highly specific conductance curves for different ions¹⁴) and receptor protein in the neuronal membrane. They would need to mimic not only the particular ion specificity of the original channels, but also their widely varying gating capacities, including non-linear voltage sensitivity (the occurrence of which in natural sodium channels is responsible for neural spiking) and ligand sensitivity (in which their conductance is modulated by signaling molecules).

We also have to deal with instances of diffuse signaling, which are not limited to a discrete and specifiable number of synaptic locations on the cell's membrane. One kind of signaling involves nitric oxide (NO), which is one of several short-lived and highly reactive molecules synthesized on demand by endothelial cells in the brain's blood vessels, as well as by astrocytes and neurons. Post-synaptic neurons use it to signal to pre-synaptic neurons (one of the mechanisms for depression and facilitation of synaptic transmission). Nitric oxide diffuses freely across cell membranes.¹⁵ How does a piece of silicon, not naturally sensitive to nitric oxide, either produce this molecule or respond to it?

More surface transducers (which could not themselves be made of silicon) would not suffice: because NO can drift through the cell membrane, the silicon replacement would need to be able to detect NO at all and any of the locations that NO could have

¹³ Giving up that anchoring raises familiar triviality problems with functionalism—see e.g. the problem of inputs and outputs, Block (1978). Perhaps these troubles can be set aside, so long as the transduction mapping from the environment to input (and from output to physical behavior) is relatively straightforward.

¹⁴ The flow of ions across the cell membrane, via these channels, is what controls the propagation (or failure to propagate) of action potentials through neurons.

¹⁵ See e.g. Hardingham and Fox (2006), Garthwaite et al. (2006), Steinert et al. (2008), and Steinert et al. (2011) for descriptions of the effects of nitric oxide in the brain. For a review of other extrasynaptic and diffusible neurotransmitters, see Del-Bel and De-Miguel (2018).

encountered in the original neuron, including internal sites where it acts to change the configuration of proteins involved in cell-internal signaling pathways—pathways that act to change future neural responses. So the silicon replacement would need to replicate the fine biophysical properties of specific proteins, at arbitrary locations within the cell, not just on its surface. This starts to look impossible, not just difficult—are there such sensors that are both small enough and able to interface effectively with a silicon chip? Or if our target neuron is itself a producer of nitric oxide, how will we store or synthesize that highly volatile stuff on demand without employing the usual enzymes and regulatory mechanisms?

c) *Biophysical sensitivities*

Our brains are sensitive to temperature, so that neurons have more trouble firing as a brain region is cooled (on a scale of 1–2 degrees C). Temperature fluctuations in our brains, perhaps caused by changes in local activity and blood flow, can have functional consequences. For a silicon chip to duplicate those dispositions to respond, it would need temperature sensors as well.

Taken together, the individual constraints surveyed so far point to a more general principle: that silicon chips are likely not to be the best substrate for an alternative realization, because responses that come naturally to chemical and biological systems have to be engineered explicitly into the silicon.¹⁶ That work is done for free in the original system—and in fact, it is probably *because* those physical features were already part of the original system that they were exploited for the system's function. I'll come back to this idea in Sect. 3, but the general idea is just the old observation that evolution tends to exploit existing features and recruit them to new uses. I think it's important not to underestimate efficiency of analog machinery compared to simulating everything digitally when there are an astronomical number of interacting components to simulate.

The system's dependence on the chemical and biological properties of its components becomes more pronounced when we consider functionally relevant activity beyond electrochemical spiking.

d) *Self-modification and other non-spiking effects*

Electrochemical activity is not the only activity exhibited by neurons that contributes to the behavioral and dispositional profile exhibited by the system as a whole. Other “outputs”—by which I mean differences that make a causal contribution to the

¹⁶ Marr (1982) mentions in passing that “some styles of algorithm will suit some physical substrates better than others. For example, in conventional digital computers, the number of connections is comparable to the number of gates, while in a brain, the number of connections is much larger ($\times 10^4$) than the number of nerve cells. The underlying reason is that wires are rather cheap in biological architecture ... [whereas] in conventional technology, wire laying is more or less restricted to two dimensions, which quite severely restricts the scope for using parallel techniques and algorithms; the same operations are often better carried out serially.” (p.24 *Vision*). As it turns out, wire is not that cheap even in biology – some of the brain's larger-scale functional organization (e.g the layout of visual cortices) seem to result in part from minimizing wiring length.

functional profile of the system—include self-modification of the cellular architecture in ways that can create new synaptic connections while eliminating old ones, or changing transcription and translation settings so that different proteins (including ion channels and neurotransmitter receptors) are present and active.

Nitric oxide—mentioned earlier—is synthesized on demand in response to various kinds of subthreshold activity. Nitric oxide can trigger internal changes that affect neural plasticity. Other diffusible neurotransmitters include endocannabinoids, carbon monoxide, and hydrogen sulfide and hydrogen peroxide.

So we have some examples of inputs that produce no immediate spiking activity which may nonetheless change the neuron in such a way that its response profile to future inputs is very different.¹⁷ In order to duplicate the behavioral profile of a neuron, the chip would need a way to store these changes so as to appropriately vary its dispositions to respond. By contrast, the biological neuron doesn't need an explicit record—the “record” effectively exists in the carefully orchestrated activities of its panoply of internal signaling molecules, engaged in finely-tuned biochemical processes, each with its own distinctive time scale relative to the input-output activities of the cell.

e) *The functional role of glia and other non-neuronal cells*

There are also biochemical functional interactions on both the input and output sides with non-neuronal brain cells (glia) which do not themselves produce action potentials. We know that astrocytes play important roles in the regulation of synaptic transmission, partly through their re-uptake of neurotransmitters at the synapse, and partly through their own release of neurotransmitters.¹⁸ As we'll see, the replacement theorist reasonably agrees that if glial cells play a non-trivial role, they will also be replaced as functional components of the system.

An astrocyte replacement would look very much like the silicon neuron replacement we have already described. It would have to be able to contact the hundreds of neurons (or neuron-replacements) within a normal astrocyte's 200–300 micron range. It would have to be able to send its modulatory packages of transmitters at the right times, in the right places. These are daunting challenges, given the tiny space we have to work with, and the material constraints imposed by the necessity of having the replacement be sufficiently physically compliant and immunologically inert so as to not damage its neighbors. Inflammation would be a major problem for any replacement that was unable to soothe the immune system.

¹⁷ Is this capacity for plasticity merely a second-order feature that doesn't matter for the core case of say, perceiving or acting at an instant? Perhaps if we are only concerned with functional duplication at the scale of hundreds of milliseconds, we may ignore plasticity, which plays a more prominent role in learning and memory. However, even at shorter time scales, sculpting or self-modification processes may matter for phenomena such as adaptation, as well as for the stability and robustness of the system.

¹⁸ See Nedergaard et al (2006), Haydon and Carmignoto (2006), Halassa and Haydon (2010) for reviews of the functional roles of astrocytes in synaptic regulation, and Allen & Eroglu (2017) for a comprehensive review of astrocyte-neural interactions. See Angulo et al. (2004) and Perea & Araque (2007) for some suggestive results about the effects of astrocytic glutamate on neural activity in hippocampal slices. Astrocytes have more recently been shown to directly control behaviors *in vivo* including chemotaxis and startle responses in fruit flies. (Ma et al. 2016)

We might be tempted to say that glia contribute to mental states *only* through their modulation of neurons. That would mean that there is after all no need to include astrocytes themselves—all we would need to do is build into the neuron-replacements the same patterns of activation that they would have had if there *had* been astrocytes. Such an attitude leads inexorably to behaviorism.

Consider: what does any particular central neuron do? Unless it is among the extreme minority that receive direct input from the periphery, or project out of the brain, what it does is modulate other neurons, its own physiology, and perhaps its own morphology. The neurons which receive its outputs then do the same. In essence, only the “output” layer, or the set of neurons immediately prior to muscle activation, can be said to be doing anything beyond modulation. The argument against including astrocytes in our alternative realization works just as well against neurons—which is to say, too well; we’d be throwing out the brain with the plumbing if we only considered the most superficial input and output layer of neurons as functionally relevant.

More generally, for the system to transition from state to state via a different process is for it to exhibit a *difference* in functional organization. If the system reaches its current state B as a consequence of the actions of glia *and* neurons in the immediately preceding state, then it is exhibiting a different functional organization than that of a system which reaches the identical state B through the actions of neurons alone, at least if the functional profile we are trying to duplicate is given at the single cell level.

Thus the temptation to subsume inconvenient functional components into their more convenient neighbors (although as we have seen, neurons are not so convenient either) should be resisted if what we are after is a duplicate that preserves the *functional organization* of the system. It’s true in any modular organization that we can shift the functions of some components into others—but then we have *changed* the organization. And if the internal organization doesn’t matter, then we are back to a kind of behaviorism—and that is a view most proponents of multiple realizability would not endorse.

2.3 Wholesale replacement and the pervasiveness of interface problems

Now it might seem that all these nasty complications arise only when we try to replace a single cell. For the purposes of an argument for multiple realizability, what really matters is the wholesale replacement scenario, in which all these messy interfaces between cells could be eliminated in favor of clean and efficient signaling connections of whatever kind turn out to be most easily implemented in silicon. Instead of changing the physical shape of cells and whether or not they physically contact each other, or changing their protein configurations in order to modulate their sensitivity, couldn’t we simply insert updating functions that tell them which inputs to pay attention to, and how much? This brings us to the next part of the replacement scenario, where, as Chalmers describes it:

“...we replace two neighboring neurons with silicon chips. This is just as in the previous case, but once both neurons are replaced we can eliminate the intermediary, dispensing with the awkward transducers and effectors that mediate the connection between the chips and replacing it with a standard digital connec-

tion. Later cases proceed in a similar fashion, with larger and larger groups of neighboring neurons replaced by silicon chips. Within these groups, biochemical mechanisms have been dispensed with entirely, except at the periphery. In the final case, every neuron in the system has been replaced by a chip, and there are no biochemical mechanisms playing an essential role. (I abstract away here from detailed issues concerning whether, for instance, glial cells play a non-trivial role; if they do, they will be components of the appropriate functional organization, and will be replaced also.)”

Now, even though each functional component still has to play the same role in the system as the original cell that was replaced, the medium in which they interact with each other is no longer constrained. Once we are looking at wholesale replacement, it seems that it should be open to us to realize the inputs and outputs of each unit in whatever form will get the job done. Surely it is contingent that our cells use *glutamate* as inputs, or *sodium* ions to produce action potentials when what *really* matters is something like “activation according to a spatio-temporal distribution characterized by this mathematical expression with these parameters”, and perhaps “activity-dependent changes to the activation rule.”

If getting an input-output-and-self-modification rule for the unit is all we need, then we should be able to dispense with at least some of the transducers we had previously been using (for intra-brain interactions), even if we still must keep the ones involved in interfacing with the rest of the body. So even though the whole system is still anchored to biochemistry at every border, we have won some elbow room.

Or have we? How many biochemical mechanisms can we really dispense with? I say it is many fewer than the replacement theorist might have thought: we will still need sensors, because in order to produce the appropriate responses to characteristics of the local non-neuronal environment, the relevant characteristics must first be detected. For example, we need to know what the concentrations of various molecules are at different locations on the interface between our wholesale replacement and the biochemistry at the borders. And we still need transducers capable of generating biochemical output if we are to maintain an indistinguishable interface with the tendrils of the rest of the biological body that are everywhere in the brain, from the vascular system to the cerebrospinal fluid.

The third main population of cells in the brain after glia and neurons are the cells associated with the cerebral vasculature. In addition to transporting metabolic fuels, oxygen, and removing waste products, blood flow is also responsible for transporting signaling molecules from elsewhere in the body, including small molecules that cross the blood-brain barrier (or else psychotropic drugs taken orally would not work). These molecules include ones that signal satiety from the gut, feedback from skeletal muscles, and hormones, which affect cognition as well as behavior.

If part of the generation of fear involves one part of the brain communicating with another part of the brain via a signal that stimulates the adrenal glands atop the kidneys to start pumping out adrenaline, then any principled account should include the adrenal glands in an organizational structure that gives an account of how the mind comes to be in a fear-state. That brain-external epinephrine source is also responsible for improved memory consolidation after stressful experiences (McIntyre &

Roozendaal 2007). A move to exclude the adrenal glands looks arbitrary from the point of view of an accurate functional characterization of the system.

The key point is that it is not enough for the replacement theorist for a replacement brain merely to produce the same outputs given the same inputs, if it does so using internal parts that have a *different* I/O profile from the original components of the system. Because we require a fine-grained *functional* duplicate, and not just a behavioral one, producing the same outputs via parts that have a different functional organization is not a genuine case of multiple realization.

The output of the blood vessels, to which every cell in the brain is exposed, makes it so that there is no clear boundary to the brain within which we can rest assured that the only inputs a brain cell receives are from other brain cells. The blood vessels carry causal influences to and from the rest of the body in ways that affect the inner functional organization of the brain, and not just its sensory inputs and behavioral outputs.¹⁹ Can we replace the blood vessels as well? Not without re-jiggering the rest of the circulatory system and the body.

The problem is that “the periphery” is actually everywhere, and interactions between the brain and the rest of the body are not neatly confined to small and localized interfaces (say at the spinal cord, cochlea, and retina). So even “just” the biochemical mechanisms at the periphery will still present a significant challenge. To make neural replacement plausible, we need to be able to partition off the parts important for mental functioning from the rest of the body. But *contra* armchair neural replacement enthusiasts, no matter how many neurons (and glia) we replace, there will always remain significant biochemical residue, because the functionally important parts of the brain are inextricably bound up with rest of the body. As Dennett puts it in *Kinds of Minds* (1996), “one cannot tear me from my body, leaving a nice clean edge.”²⁰

But even if we can’t tear the brain from the body, we might ask: why not an alternate realization of the whole body-brain bundle? The response is the same: Even if it were possible for such a bundle to realize the same behaviors, it would not be due to its *identical* functional organization. Rather, it will be because *differences* in its functional organization *accommodate* the differences in its material constitution, with mutual compensation allowing the system to approximate the original behavior.²¹ These compensatory differences, rather than making an alternative functional organization possible, instead rule it out: by the standards of the non-behaviorist functionalism we started out with, differences in functional organization entail differences in mental properties, including potentially differences in the qualitative features experienced. This is not a problem for functionalism *per se*, but only a problem for consequences of functionalism that rely on or entail multiple realizability.

¹⁹ For an exploration of the implications of these influences, see Moore and Cao (2008).

²⁰ See also Wimsatt (2006) on the shortcomings of *interface determinism* – assuming “that all that counts in analyzing the nature and behavior of a system is what comes or goes across the system-environment interface.”

²¹ Aizawa and Gillett (2009) have argued that compensatory adjustments suffice to support multiple realizability. See also Aizawa (2013).

There is also the background issue of how to circumscribe functional decompositions of the system so as to rule out gerrymandered solutions. Searle (1990) and Putnam (1988) worried that any system with sufficiently many discernible states could be interpreted as having practically any functional organization you liked.²² When there is no commitment to particular components, (cf. Godfrey-Smith (2008, 2009), Wimsatt (2007)) there is a worry that a functional description (e.g. formalized as a finite state automaton) is constrained *only* by the behavior of the whole system. As noted at the beginning of the paper, this allows for multiple realizability, but at the expense of giving any explanation of *how* mental states work.

Chalmers (1996) develops a formalism for combinatorial-state automata to address the problem and proposes that one possible restriction on functional decomposition is that the components have to be spatiotemporally continuous and relatively localized. In a similar vein, Godfrey-Smith (2008, 2009) suggests that the “obvious way to restore the lost content to functional description is to move back towards the requirement that occupants of roles have *independent standing as real parts of the system*.” So an additional reason to embrace the neural level is that those are *very* solid entities upon which to base an observer-independent decomposition of the system into functional parts. Brain cells are not gerrymandered constructs, or visible only relative to one functional decomposition; rather, they are robustly re-identifiable entities with independent theoretical and empirical justification. If we are to break up the system into different pieces, they too must satisfy the criteria for being legitimate role-occupants.

The trade-off between producing an effective behavioral profile and preserving the same functional organization is evident in existing systems that predict or reproduce very specific human behaviors, for example chess-playing machines. In many situations, a cheap chess-playing machine might produce the same moves as an excellent human player. But even at quite abstract levels of description, such a machine does not have the same functional organization as the physical realization of *our* chess-playing capacity.²³ For example, the machine will scan through many more possibilities than a human player would—it is trading off speed of search (made possible by its physical architecture) to compensate for the lack of certain kinds of human perceptual capacities (made possible by our physical architecture) that pick out more relevant possibilities from less relevant ones.

What we are supposed to be after in a system endowed with *mental states* (rather than merely artificial intelligence) is not just problem-solving, but problem-solving in a *characteristically human way*. So when the system makes errors, we want them to be the kinds of errors that *we* would make. To get that profile, we need the rest of the system, which is exactly what I’ve tried to show is extremely difficult to get without preserving the material substrate. I suppose that if we care only about states that are vaguely similar to our human ones, then a sufficiently relaxed standard of

²² See Sprevak (2018) for a thorough discussion of triviality arguments in this vein.

²³ To say otherwise would be to argue for the existence of a kind of functional structure that could not in principle be denied, regardless of the actual physical organization of the components and how they interact. That would be to leave behind the crucial functionalist requirement that functional components are individuated by their *causal* roles within the organization of the system. I will come back to this in Sect. 4.

similarity would be satisfiable even given everything that I've said. But I believe to make this move is to give up on any explanatorily interesting functionalist thesis. I'll discuss this further in the final section.

3 Generative Entrenchment and the Multiplication of Constraints

The importance of biological details described so far goes against the received view in the philosophy of mind. But perhaps this shouldn't be surprising, but rather, exactly what we should expect, given that neurons and brains are biological entities evolved within a biological organism. Complex integrated functions are less likely to be multiple realized than simple modular ones, and the principle of *generative entrenchment* tells us that we should expect complex integrated functions in evolved systems.

3.1 Multiplexed functions multiply constraints

From an everyday perspective, it may seem obvious that function constrains material make-up. A bicycle chain cannot be made of just *anything*—certain physical properties are required in order for it to perform as required within the functional organization of the bicycle. Swiss cheese would make for a terrible bicycle chain, let alone a mind. Functional properties are not nomologically independent of material properties. Furthermore, not all physical materials are candidates for all functions. There are excellent reasons that our computer chips are made of an exact combination of silicon doped with other elements, and not, say, water with a bit of CO₂. We *could* make some kind of computer out of water and pipes, but it would not have the same fine-grained functional profile at the level of speed, size, or reliability.²⁴

The usual examples used to illustrate the independence of physical description from functional description tend to be relatively simple man-made artifacts, such as mousetraps and corkscrews. These artifacts, unlike the brain, do have easily decomposable non-holistic structure. That is, there is both a clear functional decomposition of the task (bottle opening) into subtasks (grasping the cork, transmitting external force to lift the cork relative to the bottle), and a clear assignment of these subtasks to specific physical subcomponents of the artifact (the screw, the lever arms). Moreover the interface of the physical components with each other is simple (low-bandwidth), and only depends on a few of the physical properties of the components. This modular organization makes it easy to imagine replacing any one component with another one that does the same thing (relative to the relevant level of description)—“has the same input/output profile”—but is made of a different substrate.

Even the more sophisticated examples of multiple realization involve domain-specific activities. If we consider the achievements of expert systems such pokerbots, Watson, deep neural networks for image recognition, or self-driving cars, it seems plausible that highly domain-specific capacities *are* multiply realizable. But plucking

²⁴ This kind of point has also been made at length by Shapiro (2004), and Polger and Shapiro (2016). Where I have focused on the brain itself, Shapiro makes the argument in the context of mental function constraining the physical properties of the body as a whole.

out one process and showing *it* to be multiply realizable in no way shows the whole system to be so—because a single capacity in isolation doesn't have to play nice with other processes in highly complicated and specific ways.²⁵

Furthermore, the more *different* functions a particular component needs to subserve, the more constraints its make-up will need to satisfy. So for example, if a component in a functional system needs to be able to switch from high to low voltage at 100 Hz, *and* not heat up more than a degree while doing so, *and* also be sensitive to small vibrations in the table (perhaps it needs to turn off when someone walks into the room), then we are down to a very few materials, if any.²⁶

Material properties do not just come in any combinations we want—sometimes the only way to get something of a certain size and charge is also to get something gold and shiny and of a certain density. If we want an alternative realization, then even if the functional properties are preserved, some of the non-functional properties might change. If many physical properties are involved in producing the function (or there are many functions which need to be subserved, which each rely on different properties of the substrate), then there will be fewer (or no) non-functional properties left to swap out in an alternative substrate. The more different functions need to be realized simultaneously, the more bumps under the rug we will have to push around. Only the original rug will lie flat on the original fine-grained functional specification.

The functions in the electromechanical example above are *decomposable*: we could make a different subcomponent to deal with each requirement, and then put them together to satisfy the functional characterization of the part as a whole. So we could have a switching part, a cooling or thermoregulatory part, and a vibration sensor hooked up to a kill-switch. Each of those parts could be made of any number of materials (especially if the size constraints aren't too onerous), and so the whole component could also be made out of any number of materials.

By contrast, functional components of the brain tend to have functions that are *not* easily decomposable—each job is intertwined with many others, and they all need to run simultaneously. I will call this *multiplexing* of functions. Without decomposability, all the relevant functional constraints have to be satisfied by the same component at the same time, and this is the strongest reason that the brain's functional organization cannot be realized in an alternative substrate, and a constraint that holds not just for fine-scale neural function, but also more coarse-grained characterizations of the internal workings than the ones I've considered so far.

²⁵ It is sometimes argued that paradigmatic mental processes such as problem solving must be multiply realizable because an answer is an answer—whether it is spoken or displayed on a screen or demonstrated in action. But this is to confuse the medium independence of the *outputs* of a process with the multiple realizability of the process itself. Furthermore, the outputs are only medium independent because *we*, the consumers of the outputs or answers, are able to easily translate them into the appropriate form to make sense of them in the context in which they are produced. So for example, we don't care what form the outputs of the chess machine are (if they are a certain pattern of bright pixels in a display, or a voice saying "Knight to Q4", or the physical movement of a chess piece), because we are capable of translating them into the form appropriate to the exercise (of playing *chess*) that is being undertaken.

²⁶ This principle can be seen at work elsewhere in biology, for example, in the context of neutral processes in evolution: "Important proteins are more constrained and their amino acid changes are less likely to be neutral." (Ohta 2001).

3.2 Generative Entrenchment

Whence this difference between brains and artifacts? The source of the brain's entangled internal organization is what Wimsatt calls *generative entrenchment*:²⁷

A deeply generatively entrenched feature of a structure is one that has many other things depending on it because it has played a role in generating them ... Systems shaped by selection processes show significant degrees of adaptation, so things that become entrenched in them are commonly elements or parts of functional designs ... with accumulating dependencies, seemingly arbitrary contingencies can become profoundly necessary, acting as generative structural elements for other contingencies added later.

(pp 133–135, *Re-Engineering Philosophy for Limited Beings*, 2007)

As systems evolve and become more complex over time, many functions tend to get multiplexed onto the same components, because that infrastructure is already there and available—in some sense, “free.” So components that previously played only supportive or metabolic functions in an organism might become involved over time in realizing functions that seem more paradigmatically mental—for example, producing coordinated responses to a changing environment. And because the metabolic functions are crucial to the survival of the system, they cannot be eliminated without bringing the whole edifice down with them.

As the discussion on neural replacement above shows, brain function relies on a larger subset of the material properties of its realizers than human artifacts such as mousetraps. The more material properties a system exploits in order to perform its function, the harder it will be to replace the material the system is made of. Over the course of the evolution of the very complex system that is the brain, readily available properties have had their effects co-opted as functions which then contribute to the operation of the system as a whole. So while a mousetrap has plenty of “unemployed” properties to spare after its behavior and functional organization have been accounted for—properties that could be different in an alternate realization without affecting the function—the brain does not.²⁸

Moreover, the specification of a particular substrate even at the sub-cellular level produces constraints that ramify upward to the cellular level. Once a few simple molecules or components are fixed (say water or oxygen), they then constrain the components that can smoothly interface with them in the right ways, which pose their own constraints in turn on other components, until possible realizations of the whole system have been clamped down to options not significantly different from our own biological one. Being embedded in a biological system locks down the kinds of roles that functional components must play. To quote Dennett again,

²⁷ Introduced in Wimsatt (1986).

²⁸ The brain's grayish-pinkish hue is not one of its functional properties. But there is plausibly no collection of substrates, which together look green, which could also duplicate the *exact* functional properties of the brain, and so a green functional duplicate of the brain is not nomologically possible.

As systems become perceptually richer and behaviorally more versatile, it becomes harder and harder to make substitutions in the actual links of the system to the world without changing the organization of the system itself ... There comes to be a growing two-way constraint of growing specificity between the device and the environment...*We* are so multifariously and intricately connected to the world that almost no substitution is possible—though it is clearly imaginable in a thought experiment.”²⁹

Dennett’s point holds not just for the intricate connections between creatures and the world they live in, but also those between the brains, metabolic systems, and bodies of creatures with minds. Metabolic stuff is tightly intertwined with the information-processing stuff, and in the brains of creatures like us, you cannot have one without the other. I suspect that few proponents of MR would want to argue that metabolism is multiply realizable in the same way—perhaps because the *point* of metabolism is to exploit particular substrate-specific physical properties (e.g. of storing or releasing energy), rather than abstracting from them so as to implement some informational process. But if metabolism is tied to a particular set of substrates, then it will anchor the information-processing stuff along with it.

3.3 Examples of generative entrenchment in the brain

Glia are a good example of the evolutionary co-option of existing infrastructure for novel functions. Glia precede neurons in development; they are necessary for neural survival in virtue of providing metabolic support, and they modulate neural activity responsible for cognitive activities such as memory formation. More interestingly, there is some evidence (e.g. Han et al., 2013) that glia play a more substantive cognitive role in humans than in simpler organisms (which don’t exhibit those cognitive capacities in the first place). So we see a component that initially plays only a supportive or permissive metabolic role becoming integrated into the cognitive architecture over evolutionary time. This pattern of dependence between metabolism and what might be termed “informational” capacities shows up at many different scales within the organism.

At the cellular level, neural adaptation is a metabolic feature of neurons that also plays an information processing role. In adaptation, cells quickly become less excitable after initial activation. This can be a direct consequence of resource depletion—it takes time for a cell to recover completely after firing an action potential, or for a synapse to “recharge” after releasing neurotransmitters. But in sophisticated nervous systems, adaptation has come to be exploited by the system as a way of detecting novelty—a change in the patterns of stimulation produces more activation than the same stimulation over and over again. Given a framework in which lowered activity functions to signal lack-of-novelty, further mechanisms can then be scaffolded on, in which active suppression (perhaps top-down) can play the same role of downgrad-

²⁹ Dennett (1990). My contention is that the “thought experiment” proposed is one that must ignore nomological constraints such as chemistry-in-our-world. Although Dennett is often considered a behaviorist, his arguments lend a surprising amount of aid and comfort to this chauvinistic functionalist.

ing the significance of some response for the rest of the system. When fMRI studies invoke “repetition suppression” now, they use a name which highlights the function as an active part of the system, rather than merely a side effect of temporary resource depletion.

Zooming in to the molecular level, we have ATP (adenosine triphosphate), commonly described as the energy currency of the cell. ATP evolved long before any kind of nervous system. In the brain, as elsewhere in the body, ATP is the high energy product of cellular respiration, and is trafficked around the cell and used as the mobile power source for ion pumps, protein regulation, and protein synthesis. These metabolic functions appear entirely and merely/only biological, and one might suppose that they play no role in causal economy of the brain that gives rise to mental states. However, ATP also plays a role in intercellular signaling in the brain. Astrocytic calcium waves, which have been shown to modulate neural activity, propagate via ATP. ATP is also degraded in the extracellular space into adenosine, which inhibits neural spiking activity, and seems to be a primary determinant of sleep pressure in mammals. Within neurons, ATP hydrolysis is one of the sources of cyclic AMP, which is involved in numerous signaling pathways within the cell, including those that modulate the effects of several neurotransmitters producing effects on the neuronal response to subsequent stimuli in less than a second. Cyclic AMP is also involved in the regulation of intracellular calcium concentrations from internal stores, which can have immediate and large effects on the efficacy of neurotransmitter release from a pre-synaptic neuron.³⁰

We can see that the kind of physical realization that whatever fulfills the role of ATP in the brain has to have is strongly constrained by the multiple roles it needs to play. While ATP’s close chemical and biological cousin GTP (guanosine triphosphate) can and does also work as an energy supply in the cell, it cannot perform the additional suppressive modulatory effects of extracellular ATP on neurons, because neurons have adenosine receptors, not guanosine receptors, to subserve those regulatory functions.³¹ Furthermore, since the relative concentrations of GTP and GDP are an important determinant of the speed and efficiency of G-protein coupled second messenger signaling cascades within the cell, having extra of either floating around would have a significant effect on the cell’s normal transactions, disrupting some and introducing others.

ATP is not alone in playing many roles. Lactate, an end product of metabolism in the brain, also serves as a signaling molecule that regulates neural excitability and plasticity (Magistretti & Allaman, 2018). BDNF (brain-derived neurotrophic factor) and other growth factors (including those released by blood vessels), products of the endocrine system, and classical neurotransmitters such as dopamine also play a dizzying array of simultaneous roles in the brain and body, both in development and in adulthood. Elaborate feedback regulation control ensures that the different roles are

³⁰ See Greengard et al (2001) for a review of second messenger systems in slow synaptic signaling, and Abbracchio et al. (2009) for purinergic (e.g. ATP) signaling in neurons, and Hamilton & Attwell (2010) and Zhang et al. (2003) on purinergic signaling from astrocytes.

³¹ Of course, neurons also have guanosine-sensitive receptors, but those produce yet *other* modulatory effects, some but not all of which overlap with the effects of adenosine.

balanced against each other, but also produces a network of interdependencies so that perturbations in any of the systems in which the molecule is involved will produce further disturbances in the others.³²

Molecular-level constraints can have striking consequences at the psychological and behavioral level. Dopamine, for example, is involved in generating motor commands, reward prediction and learning, and regulation of blood flow. Fiddling with one function perturbs them all. For example, when patients with Parkinson's disease are given L-DOPA (a dopamine precursor), a significant fraction develop gambling disorders even as their motor symptoms are relieved.³³

Because of shared substrates, not just metabolism, but motor control, reward-based learning and other cognitive control functions become tied together in the functional architecture. Given a brain in which even simple molecules have many different roles to play (with effects at many size and time scales), and given the nature of the system, in which small differences can be amplified to have large effects, the constraints can quickly become prohibitive.³⁴

3.4 Why can't multiplexed functions be unbundled and separately realized?

Why is it that the multiple functional roles played by particular components cannot be decomposed, allowing individual functional roles to be alternatively realized by separate components? Why can't we separate out the energetic role of ATP, and each of its distinct signaling roles from each other? Why not have one component that plays the energetic role, another that plays the extracellular signaling role, a third that plays the intracellular modulation role, and so forth? Suppose we did, along with further components that make sure that these functions are all appropriately coupled, so that when ATP is depleted, they all slow down (i.e. preserving the original functional profile of the component). What then prevents us from putting a black box around the whole ensemble and saying, *this* is a single functional component that plays the ATP-role in the system.

Such an approach may work if there are not too many components or too many functions per component. But when there are many components and functions that would cross-cut the compositions of individual black-boxes, then attempting to duplicate functional interactions among them would result in a combinatorial explosion of further components to discipline those interactions in the same way that they are constrained by their substrates in the original system. If on top of this there are space, timing or robustness requirements on the behavioral profile of the system, as we would expect for a functional duplicate of a mind, then the requirements will be difficult to satisfy.

³² Krishnan and Nestler (2010) provide an accessible review (focusing on depression) of some of these factors.

³³ Pirritano et al.(2014); Schultz et al. (1997); for blood flow, see Krimer et al. (1998); for a review of the role of dopamine in cognitive control, see Ott and Nieder (2019).

³⁴ Another term from Wimsatt (2007), the "causal thicket", seems appropriate here. Causal thickets are where distinct levels and perspectives break down, so that no clear size scale or methodological foregrounding seems appropriate: "The neurophysiological, psychological, and social realms are mostly thickets." (p.239).

Perhaps someone who believes that mental properties are *inherently* coarse-grained compared to most salient biological properties will want to get off the boat here—even if modular decomposition of the system fails at the cellular or molecular level, it might still in principle re-emerge at a higher functional level. But it turns out that brain functions exhibit this lack of modularity at higher levels of description as well, meaning levels of description that might focus on cognitive or perceptual capacities rather than neurophysiology, as well as levels of description that focus on brain areas, rather than molecules or cells. So just as there isn't only one thing that nitric oxide does, that we could draw a box around, and replace the whole nitric-oxide system with a functional duplicate of, or a single thing that glutamate does, or only one particular thing that interneurons do, so too with larger chunks of the brain, characterized by what they do, such as "the visual system", or where they are, such as "the pre-frontal cortex". Parts of our visual systems are not just for figuring out what and where things are, but also contribute to the coordination of our circadian rhythms with external zeitgebers, which in turn regulate patterns of gene expression throughout the body, including elsewhere in the brain in ways that dramatically affect behavior. Going the other direction, if we consider a particular capacity, it often involves the interplay of many different brain parts, with important processes happening at many different temporal and spatial scales. Consider visual contour detection. The process involves the retina, on a very local scale, as well V1 and other brain regions along the visual pathway. Our absurdly good ability to detect even very noisy, faint, or fleeting contours is not based on acuity alone, but some complicated combination of "knowing how things ought to look" based on high-level (perhaps conscious) priors, and various parts of the visual system putting edge detection together with color discrimination, along with other low-level heuristics that we have no conscious access to. There is no single "contour detection" box that could be replaced in isolation without interfering with the other functions that share infrastructure with the processes realize contour detection. The mapping from psychological capacities (or brain functions that we take to be important) to physically individuated parts in the taxonomy of neuroscience is in general not one-to-one.³⁵

4 Can we loosen the constraints on multiple realization?

We've now reviewed some specific empirical facts that militate against the plausibility of neural replacement. I've argued that the principle of generative entrenchment can help us to see why it is not surprising that the functional causal roles in a complex evolved system will be interwoven so as to produce many constraints in their possible realizers. But it may seem that the argument proves too much. If the relevant "behavioral" profiles are specified by maximally fine-grained response properties in *every possible* context, then it might turn out that *nothing* is multiply realizable: if there is *any* physical test that could distinguish the original from its putative duplicate, then they are not functionally identical with respect to their responses to that test.

³⁵ See Poldrack (2006) and Anderson (2010) for summaries of the evidence against such proprietary assignments.

I agree that fine-grained neural functional duplication is too demanding. That was only a convenient starting point, convenient both because the MR claim at the neural level has seemed so plausible to so many, and because we think that (all else equal, to account for meaning externalism) duplication of neural functional profiles should be sufficient for duplication of mental properties, given the usual physicalist assumptions about supervenience.³⁶ But when we find that duplicating *all* the lower-level processes that have some effect on behavior turns out to be overwhelmingly difficult, and in retrospect unsurprisingly so, then we should pull back and commit only to what we really need, which is: functional duplication at just the level of detail required to preserve the relevant *psychological* properties, but not all functional properties whatsoever.

Standard functionalism, of the kind adopted by cognitive scientists and more recent philosophers, presumes some intermediate level of abstraction, below behavior and above biology, that is “the functional level”. (Compare: Marr’s (1982) representational or algorithmic level). It is on this basis that functionalism has gained most of its wider popularity in cognitive science as a foundational truth that underlies the plausibility of artificial intelligence and computational explanation more generally.

But we have no theoretical guarantee that any such efficient, intermediate level of description exists; perhaps there are no real patterns at the sub-personal level that are well above the level of biological description. Or perhaps entities and organized patterns at such a level of abstraction do exist, but then the burden of proof is on the functionalist to show that they do, and to characterize them in empirically relevant (i.e. non-trivial) ways. By this I mean, such entities and patterns should be independently identifiable, not just imputed from behavior, but explicitly measurable using the tools of neuroscience.

It also seems likely that different abstractions may turn out to be appropriate and effective for characterizing different psychological phenomena—and so there would not be a single functional level of abstraction, but many. It might be too optimistic to suppose that that very same abstraction would help us to explain everything else that we think matters for having a mind, especially if we can’t point to any mechanism that would enforce that level of abstraction the way that we do in our built computational machines. Until cognitive science or neuroscience provide evidence (and not just vague hopes) for the utility of understanding the system in this way, the question remains empirically open.

In the meantime, there are two practical ways to think about more realistic replacement scenarios. One way is to characterize the functional profiles (including behavioral outputs) that need to be duplicated more coarsely, ignoring insignificant differences in behaviors. Perhaps we don’t care if an android speaks in a monotone, or can’t use contractions, if it nonetheless behaves in a human-like way in broad strokes. But now consider the level of description at which a notebook and our internal biological memory count as playing the same functional role. Plausibly that level

³⁶ As Lycan (1981) writes, “‘Neuron’ can be as much a functional description as ‘pain’. It is just empirically implausible that mental states are best identified with something so specific as the neuroanatomical level.” And so one consequence of a very fine-grained functionalism is that it pushes us towards a biologically chauvinistic theory of the mental.

of description is so coarse that we are no longer able to make psychologically interesting distinctions. So we have to navigate carefully here: functionalism can throw some details away, but the relevant level of description has to retain enough detail to *explain* (and not just re-describe) interesting psychological and behavioral phenomena, distinguishing between states and processes that are of distinct significance in our mental lives.

A natural level of abstraction at which to characterize the functional profile comes from the more scientifically pragmatic question: what is the functional organization that characterizes a *typical* brain—and is the duplication of that functional organization sufficient to reproduce whatever psychological features (themselves specified to a pragmatically reasonable degree of precision) are exhibited by creatures with those brains?³⁷ That means that the level of detail we are committed to duplicating is just that level at which we can plausibly say that different individuals within our group of interest count as having the *same* functional organization, in producing behaviors at the level of detail that we are interested in differentiating within. As we know, there is substantial diversity among humans in both brain architecture and in functional profile. What we all have in common is a typical human psychological profile realized by a typical adult human brain. Those typical functional states would gloss over individual differences, so that a range of exact states would qualify as instances of the same functional state. I will come back to this approach in the last section.

A second way to pull back is to characterize the functional organization itself at a coarser grain, breaking down the system into larger functional components. Now the relevant functional states will be those displayed by a group of neurons that have particular statistical or group dynamical properties over some period of time, with the presumption that those ensemble properties are the right level at which to specify the functional role of the component (which is now a group of cells that we have reason to believe tend to work together). This population coding assumption is widely accepted in contemporary neuroscience. Not only is population activity often more practically accessible, but it seems clear that groups of neurons that exhibit considerable individual variation at the cellular level (since two cells in two mammals are exactly alike) still perform very similar functions in the respective brains in which they reside.

Unfortunately, there is reason to think that silicon chips will face similar problems trying to duplicate the function of groups of neurons as they do trying to duplicate the single neuron—and in this vein the considerations sketched below should sound familiar.³⁸ Within the original biological system, regularities are maintained (despite

³⁷ Different investigations might prioritize different sub-questions, focusing on the human, or primate, or mammalian, or drosophila brain, for example. And for a few quite sophisticated capacities (as well as many simpler ones), we are now able to simulate species-typical responses within a plausible marginal of error (see e.g. Yamins and DiCarlo 2016).

³⁸ The literature on neural population coding is huge, but for an accessible review, see Panzeri et al (2015). Wei et al (2019) explore various analysis techniques to deal with the variability in neural responses, and also provide a nice discussion of the virtues and drawbacks of these approaches. To be clear, my reservations about multiple realization at the level of a typical organism's brain function apply primarily to the global collection of functional states and processes taken to constitute unfettered mental states in the wild. For extremely specific capacities to be demonstrated under relatively restrictive conditions, by contrast, I

internal variability of the components) by redundancy and fine adaptive adjustments. It is these fine adaptive adjustments that we would be eliminating by replacing a neuron with a cruder chip that behaved on average in the same way, but not in every instance in the same way. (Anyone who has learned to ride a bike, ice skate, or better, walk a slackline rope will be familiar with the importance of small dynamic adjustments. When coasting straight on a smooth surface, it looks like the rider is just holding the handlebars perfectly still, and keeping their weight unchanged—and indeed on average they are. But if you fix the handlebars, replace the person with a mannequin, and give it a push, it won't stay upright for very long). Selection for stability in the original biological system will tend to create conditions where noisy variability in individual cells is compensated by other cells in the system; that compensation would no longer be present in the replacement system.

Another way to think about this: even though variation, or perhaps even ablation, of a single neuron will have little (if any) effect on the behavior of the population, it does have an effect on the robustness of the behavior that is preserved. As more neurons are replaced, the ability of the network to compensate for variation will also decrease. When all the originals in a group have been replaced, we might have a very fragile copy of one particular set of input-output relations (rather like a thin look-up table). As computational modelers have found, it is easy to get undesired epileptic behavior in simulated neuronal networks. There, tiny changes (well within the range of variation exhibited by actual cells) in the units that make up the simulated population can have dramatic effects when the units interact with each other. So it is difficult to duplicate the robustness of the original functional profile.

What if we engineer the replacements to exhibit compensatory adaptation to neighboring activity as well? Well, consider what underlies the ability to dynamically adjust to contextual changes in environment in a local way (that is, without having to predict just what effect a single cell would have on the population as a whole in order to work back to how it should compensate for what its neighbors are doing, even as they too change in an attempt to do the same thing). One important compensatory mechanism is *synaptic scaling*, in which neurons homeostatically regulate their sensitivity to input in order to maintain a given level of activity.³⁹ This regulation is deeply embedded in the biological/metabolic function of the neuron, involving protein synthesis and intracellular transport mechanisms. Once again, generating the right set of biochemical outcomes will be complex and perhaps intractable—because of the huge numbers of degrees of freedom in simulating biochemical interactions. So unless there is an efficient rule that generates the dynamics of synaptic scaling (while ignoring biochemical interactions involved in cellular metabolism), a silicon replacement will not have the ability to compensate for variability in its neighbors.

Of course, if we were able to directly computationally specify how the population behavior evolves over time in response to inputs, we could bypass the prob-

am agnostic, or even optimistic, since many fewer functional properties will matter, and there have been striking empirical successes on this front.

³⁹ For a review of homeostatic mechanisms of activity regulation, see Turrigiano (2007). Glia also contribute to this homeostatic mechanism by release of growth factor TNF-alpha, see Stellwagen and Malenka (2006).

lem of constituent adaptation. But that problem is just as much of an obstacle at the population level as at the single cell level, because while generalizations about neural function are possible—and indeed, comprise much of our scientific understanding of neural behavior—they hold only in virtue of a vast (infinite) set of *ceteris paribus* conditions. Not only is there no way to specify every factor in the *ceteris paribus* conditions, there is reason to believe that many of them would be violated in the scenario where we have tried to eliminate as many of the biochemical mechanisms as possible.⁴⁰

The physiological considerations surveyed so far have been quite general, in the sense that they are shared across humans and other animals. (In fact much of our knowledge of neurophysiology comes from investigations in model organisms such as mice and monkeys). Thus the relevant evidence they provide against replacement applies even to systems functionally characterized at the coarser grain that puts us together with animals. In other words, nothing in my argument shows that other similar *biological* systems couldn't implement the same function; it is the difference between biological realizations (with all their evolved capacities for dynamic adaptation, including changes in structure and so forth) vs. non-biological substrates that makes the difference here.

In fact, part of the original motivation for multiple realizability comes from a sympathetic evaluation of the animal kingdom. It certainly seems as though many aspects of a basic mental profile are realized by many animals, including animals which display significant variation in gross morphology. If their bodies are unlike ours while their minds are like ours, don't they stand as actual living examples of multiple realization? I say no, because to suppose this would be to ignore the fact that at a cellular level, animal brains are very much like ours, even if their gross organization and outward appearance seem utterly alien. Putnam's octopus, for example, shares many features of our pain-related architecture, including peripheral touch receptors mediated by chemical neurotransmitters, sensory areas that are distinct from motor areas in a brain full of neurons, and endogenous opioids that act on internal receptors that are genetically similar to ours in much the way that our opioids act on our opioid receptors. So at least some apparent differences are belied by underlying similarities when we consider the relevant physical properties important for realizing the relevant functional and behavioral properties. This is not to deny that octopuses do have quite different nervous systems than we do at some levels of description, but then, they also have quite different behavior from ours, issuing from those very physical differences. Similar points can be made about plasticity and re-use within a single animal (as described e.g. in Sharma et al (2000), or Anderson (2010)). These are cases where the same brain tissue may subserve apparently quite different functions - showing that those functions don't rely on a geographically identified set of cells, but not showing

⁴⁰ Experimentalists can run afoul of these conditions when they do experiments on neurons either in culture, or in acute slices, rather than *in vivo*. For example, neurons in dissociated cultures express unusual complements of receptors that are not seen *in vivo*, because of differences in their exposure to various growth factors. Neurons in slices are much more likely to exhibit oscillatory behaviors that are rarely seen *in vivo* (leading to some contentious debates about whether the oscillatory phenomena studied are genuinely relevant to brain function). And neurons can't survive without glia, either in or out of the brain.

that what makes all (any) of the functions realizable by brain tissue could be realized without the specific chemical and biological properties of brain tissue.

To defend multiple realizability, the functionalist has to be able to separate out the parts of the causal structure that matter in the brain from the ones that don't (if any). Because of the nested and intertwined nature of processes undertaken by cells in the brain, the result of generative entrenchment, this separation seems unlikely at the cellular and molecular level. The brain is an evolved biological entity tightly integrated with the body it sits in, and as such, many of its metabolic and permissive functions are intertwined with its cognitive and mental ones. To untwine them in order to allow for alternative realizations would require changing the very functional organization that the functionalist wants to preserve.

5 Implications for Functionalism

Functionalism as originally conceived by Putnam took “functional organization” to be the organization of states of the whole system, given by the machine table of contingencies and transitions between states of the entire system, which could in theory be understood independently from the physical organization of discrete components (though of course it needn't be). It was that conception of functionalism, with token identities such that functions did not need to depend on any particular individuation of functional parts, that gave us materialism as merely a light cloak on a relatively unconstrained theory of mental states.

Machine functionalism has mostly been left behind, but many philosophers still take the spirit of functionalism to be freedom—freedom from material constraints and the messy details of biology, freedom from chauvinism and reductionist naïveté, and freedom to embrace a clean computational model of the mind. I've argued that standard functionalism, understood in terms of the causal roles played by the constituents of a complex system, is more like an iron cage—and that the kind of causal structure necessary for a mind like ours can be satisfied only by a body like ours with a brain like ours.

Everything that we know about neurons suggests that it would be very difficult for a silicon chip to replicate the function of a neuron at the finest level. If it is possible, it seems to me the burden of proof is on the functionalist to show that it is. If I'm right that fine-grained replacement is not just practically unfeasible, but actually nomologically impossible, then the *only* functionally isomorphic systems (even when individuated more coarsely) are systems made of biologically similar stuff, and we are stuck inside the biological cage. Abstract functional organization, it will turn out, is not so abstract that it allows us to ignore the material substrate—rather, it is sensitive to biological realities and the biological (physical and chemical) properties responsible for them.⁴¹

⁴¹ The importance of these constraints also provides a convenient *functionalist* argument for why Blockhead and Aunt Hilary don't qualify as having mental states—because they are misdescribed as realizing the right functional organization, even though it would in fact be nomologically impossible for them to have the right functional organization due to the constraints that function places on constitution.

This implication of functionalism cannot be what anti-chauvinists had in mind when they embraced functionalism as a replacement for the identity theory. But I think a chauvinist functionalism has much to recommend it. As Bechtel and Mundale (1999) write at the end of their repudiation of multiple realizability: “Even if one can identify psychological states with activity in brain areas, that does not render the functional characterization of psychological states any less important ... A thorough-going functionalism uses functional criteria to identify both psychological states and brain states, and can survive even if we jettison the multiple realizability thesis.”

If all we want from functionalism is a conceptual analysis of mental state talk—that *analytical* functionalism in no way threatened by the implausibility of multiple realizability.⁴² Some functionalists, including Lewis and Armstrong, and perhaps also Clark (2008) are understanding their project in this way. David Lewis (1972) outlined a project for characterizing functional roles at the personal level—to be derived from the platitudes, suitably regimented. But at that maximally coarse-grained level, functionalism says very little if anything about causal roles at any subpersonal level. And even at the top (personal) level, this kind of functionalism is only a schema, and one that has yet to be filled out. Perhaps folk psychology is rich enough to specify all the functional roles. But it is not the sort of thing that can tell us more than we already knew from the armchair.⁴³

In *The Mind Incarnate*, Shapiro writes, “Functionalism’s great promise lay in its suggestion that it is possible to investigate minds at a level of description that makes unnecessary an appreciation of the bodies and brains in which they are implemented.” What Dennett (1986) has called “High-Church Computationalism” relies heavily on this promise, in its rejection of the relevance of neuroscience to the study of the mind. But this promise depends entirely on the truth of the empirical bet of multiple realizability. Losing the bet means that functionalists have to embrace the hard work of getting to know the messy biological details, if functionalism is to serve as the foundation for a productive research program.

But let me end on an ecumenical note. Arguably what the functionalist really wants (and needs) is *not* the possibility of some *other* substrate playing the same functional roles as our biological stuff—but rather, the validation of an abstract level of description that is sufficient to generate the phenomenon of interest. If it is possible to specify the functional roles and organization at an *informative* level of detail that is nonetheless abstract enough to be tractable, then we will have specified an

⁴² It might be argued that functionalism without multiple realizability isn’t functionalism at all, because it doesn’t give us the non-reductive goods—i.e. the autonomy of psychology. I think this is a mistake: to the extent that multiple realization is an empirical claim on which autonomy depends, any theory that presumes autonomy from the armchair is exhibiting a failing and not a virtue. I don’t think that functionalism must be committed to the autonomy of psychology—rather, a functionalist, behaviorist, and identity theorist might all reach the same verdicts as to which actual systems count as having minds and psychological states, but for different *reasons*. (And because the reasons that I rely on are ultimately functionalist, the view I am defending here is not a brute identity theory.)

⁴³ This is a point raised in Dennett (1987). Maybe this explanatory difference is why functionalism and behaviorism can feel so different: behaviorism explains particular behavioral profiles by reference to the general (e.g. she yelped upon stubbing her toe because that is an instance of what people do under general circumstances of minor bodily insult), while functionalism at least purports to explain particular behavioral profiles by reference to their internal causes.

explanatory and predictive model of the phenomenon of interest. It then becomes less pressing whether we could implement the model in some *different* substrate, so long as we can query the relevant aspects of the model to engage successfully in our explanatory practices.

The scientifically helpful version of functionalism looks for a sweet spot of abstraction—above those messy details that don't matter for the phenomenon of interest, but with structure still sufficiently articulated to be more informative than behaviorism. So the real challenge for a scientifically relevant functionalism is to discover whether such a level (or levels) of abstract description exists, and then to characterize them if they do.⁴⁴

Whether such a tractable level of description exists might be expected to differ in different domains. So we might be relatively optimistic about the promise of neural prostheses and practical brain-computer interfaces *in just those cases where* we are concerned with a narrowly delineated task. This is because when the interface is low-dimensional and context-specific, many of the holistic constraints on the system are lifted. With fewer employed properties, and more leeway in producing functional differences elsewhere in the system (since we are only interested in duplicating a particular behavioral capacity), the options for compensatory adjustments are abundant. The same may also be true for neural network models of particular aspects of our mental capacities, again taken in isolation. But if I'm right and the biological details matter for maintaining complex integrated functional profiles, then we should be skeptical about the possibility of uploading one's entire cognitive economy to a virtual environment (realized in an artificial computational system) while preserving the relevant cognitive capacities.

Functionalism was originally supposed to be the best of both worlds: as materialistically respectable as the identity theory, but more informative than behaviorism, while avoiding both chauvinism and liberality. Ironically, taking functionalism seriously points us back in both directions at once: Towards a methodological behaviorism on the one hand—where we specify capacities and phenomena that we are interested in, top-down, and then ask a general computational question about what it takes to realize those capacities and phenomena; *and* towards a biologically-focused identity theory on the other—where we pay attention to the messy details of the intricate dance of complex biological processes selected over billions of years to produce just what it takes to do well in a demanding world. If the constraints that complex functions impose on their realizers are sufficiently strong, then it might even turn out that these two approaches meet in the middle—so that not only functionalism, but

⁴⁴ A series of related points are made beautifully in Sect. 4 of Chirimuuta's (2018) review of Polger and Shapiro (2016). She emphasizes the importance of functional thinking in biology itself, and the interaction of this fact with the Heraclitean nature of living matter: "It can both be true that the material from which the nervous system is built (i.e. living, metabolizing cells) is crucial to their function *and* that those functions are multiply realized... Biological systems robustly maintain their functional profiles in spite of constant internally and externally generated perturbations. Therefore, the functionally relevant patterns—which stay the same across lower level changes, like the spiral shape of a raging tornado—will not be readily apparent at the finest grain descriptions of individual cells, or even small circuits. Hence the need for "meso-level" descriptions which make salient the shape of the storm against the swirling flux of background changes." We may disagree about whether such multiply-realizable meso-scale descriptions are likely to exist, but not about their explanatory desirability.

behaviorism forces us to be biological chauvinists about what it takes to have a mind in a world like ours.⁴⁵

If we want to understand *how* the mind manages to do what it does, both of these final two approaches are valuable: To understand specific details about our minds we may need to learn many messy biological details. But to understand the general features of possible cognitive architectures, we can instead turn to whatever physical architectures work, whether they resemble us in organizational and physical aspects or not. Successes in artificial intelligence come not from trying to duplicate our own functional architecture at a very fine level, but rather from exploiting whatever substrates and functional architectures are available for the ways in which they support complex (but intelligent and intelligible) processes, different from our own from the ground up.

Acknowledgements My heartfelt thanks to Peter Godfrey-Smith, Sean Kelly, Bill Wimsatt, Daniel Dennett, Nicholas Shea, Larry Shapiro, Tom Polger, Enoch Lambert, Justin Junge, Beckett Sterner, Brian Hemond, Chris Moore, Evelyn Keller, Sam Tobin-Hochstadt, Katie Edmonds, Lindy Blackburn, Nikhil Bhatla, Rob Goldstone, Colin Allen, Peter Todd, Kirk Ludwig, Gary Ebbs, Gualtiero Piccinini, Ned Block, Dan Waxman, Jeremy Goodman, David Chalmers, Michael Strevens, Charles Rathkopf, Helen Longino, Michael Anderson, Daniel Yamins, Paul Nuyujukian, Keith Winstein, and Jared Warren for extensive and helpful comments over many revisions. I am also grateful for feedback and suggestions from audiences at BU, Tufts, IU, Antwerp, Irvine, USC, and Stanford, as well as several anonymous reviewers.

Conflict of interest

No funding or conflicts of interest.

References

- Aaronson, S. (2011). Why Philosophers Should Care About Complexity, in Copeland, B., Posy, C., and Shagrir, O. (eds.), *Computability: Turing, Gödel, Church, and Beyond*, MIT Press, (2013)
- Abbracchio, M. P., Burnstock, G., Verkhratsky, A., & Zimmermann, H. (2009). Purinergic signalling in the nervous system: an overview. *Trends Neurosci.* 2009 Jan;32(1):19–29.
- Aizawa, K. (2013). Multiple realization by compensatory differences. *Euro Jnl Phil Sci*, 3, 69–86.
- Aizawa, K., & Gillett, C. (2009). The (Multiple) Realization of Psychological and other Properties in the Sciences. *Mind & Language*, 24, 181–208.
- Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. (Target article. *Behavioral and Brain Sciences*, 33(4), 245–266.
- Angulo, M. C., Kozlov, A. S., Charpak, S., & Audinat, E. (2004). Glutamate Released from Glial Cells Synchronizes Neuronal Activity in the Hippocampus. *J. Neurosci.*, 24(31 (August 4), 6920–6927.
- Allen, Nicola J., Eroglu, C. (2017) Cell Biology of Astrocyte-Synapse Interactions. *Neuron* 96(3) 697–708 S089662731730925X. <https://doi.org/10.1016/j.neuron.2017.09.056>
- Bechtel, W. and Mundale, J. (1999). “Multiple Realizability Revisited: Linking Cognitive and Neural States,” *Philosophy of Science* 66, no. 2: 175–207.
- Bickle, J. (2010). Has the last decade of challenges to the multiple realization argument provided aid and comfort to psychoneural reductionists? *Synthese*, 177, 247–260.

⁴⁵ See Yamins and DiCarlo (2016), Kell et al. (2018), Schrimpf et al. (2021) for intriguing examples of strong task-imposed constraints on plausible architectures. These are also good examples of relatively successful alternative realizations of narrowly delineated cognitive capacities.

- Block, N. (1978). 'Troubles with Functionalism'. In C. W. Savage (Ed.), *Perception and Cognition: Issues in the Foundations of Psychology*. Minneapolis: University of Minnesota Press.
- Block, N. (1997). "Anti-Reductionism Slaps Back" *Noûs*, Vol. 31, Supplement: Philosophical Perspectives, 11, Mind, Causation, and World, pp. 107–132.
- Bullock, T. H., Bennett, M. V., Johnston, D., Josephson, R., Marder, E., & Fields, R. D. (2005). The neuron doctrine, redux. *Science*, Nov4;310(5749):791–3.
- Cao, R. (2014). Signaling in the Brain: In Search of Functional Units, *Philosophy of Science*, 81 (December 2014) pp. 891–901.
- Chalmers, D. J. (1995). Absent Qualia, Fading Qualia, Dancing Qualia. In Thomas Metzinger (ed.), *Conscious Experience*. Ferdinand Schöningh. pp. 309–328.
- Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese*, 108, 309–333.
- Chirimuuta, M. (2018). Marr, Mayr, and MR: What functionalism should now be about. *Philosophical Psychology*, 31(3), 403–418. <https://doi.org/10.1080/09515089.2017.1381679>
- Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press, 2008.
- Dennett, D. C. (1986). "The Logical Geography of Computational Approaches: A View from the East Pole." In *The Representation of Knowledge and Belief*, edited by R. Harnish and M. Brand, 59–79. Tucson: University of Arizona Press.
- Dennett, D. C. (1987). "Fast Thinking" Chapter 9 from *The Intentional Stance*. Cambridge MA: MIT Press.
- Dennett, D. C. (1990). True Believers: The intentional strategy and why it works, W. Lycan, ed., *Mind and Cognition: A Reader*, MIT Press.
- Dennett, D. (1996). *Kinds of Minds*. New York: Basic Books.
- Del-Bel, E. & De-Miguel, F. (2018) Extrasynaptic Neurotransmission Mediated by Exocytosis and Diffusive Release of Transmitter Substances. *Frontiers in Synaptic Neuroscience* 1013. <https://doi.org/10.3389/fnsyn.2018.00013>
- Fang, W. (2018). The Case for Multiple Realization in Biology, *Biology and Philosophy* 33 (1–2):3 (2018).
- Fodor, J. (1974). "Special Sciences (Or: The Disunity of Science as a Working Hypothesis. *Synthese*, 28, 97–115.
- Garthwaite, J. (2006). Signaling from blood vessels to CNS axons through nitric oxide. *J Neurosci*, 26, 7730–7740.
- Greengard, P. (2001). "The Neurobiology of Slow Synaptic Transmission", *Science* 2 November 2001: Vol. 294. no. 5544, pp. 1024–1030.
- Godfrey-Smith, P. (2008). Reduction in Real Life, In Jakob Hohwy & Jesper Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press.
- Godfrey-Smith, P. (2009). Triviality arguments against functionalism. *Philosophical Studies* 145, 273–295.
- Halassa, M., & Haydon, P. G. (March 17, 2010). Integrated brain circuits: astrocytic networks modulate neuronal activity and behavior. *Annual Review of Physiology*, 72, 335–355.
- Hamilton & Attwell Nicola. (2010). "Do astrocytes really exocytose neurotransmitters?". *Nature Reviews Neuroscience*, 11, 227–238.
- Han X, Chen M, Wang F, Windrem M, Wang S, Shanz S, Xu Q, Oberheim NA, Bekar L, Betstadt S, Silva AJ, Takano T, Goldman SA, Nedergaard M. (2013) Forebrain engraftment by human glial progenitor cells enhances synaptic plasticity and learning in adult mice. *Cell Stem Cell*. 2013 Mar 7;12(3):342–53. <https://doi.org/10.1016/j.stem.2012.12.015> PMID: 23472873; PMCID: PMC3700554.
- Hardingham, N., & Fox, K. (2006). The role of nitric oxide and GluR1 in presynaptic and postsynaptic components of neocortical potentiation. *J Neurosci*, 26, 7395–7404.
- Haydon, P. G., & Carmignoto, G. ; Astrocyte Control of Synaptic Transmission and Neurovascular Coupling. *Physiol. Rev.* 86: 1009–1031, (2006).
- Hollingsworth, S. A., & Dror, R. (2018). Molecular dynamics: Simulation for all; *Neuron* 99, pp. 1129–1143.
- Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, 98(3), 630–644e16.
- Krimer, L. S., Muly, E. C. 3rd, Williams, G. V., & Goldman-Rakic, P. S. (1998). Dopaminergic regulation of cerebral cortical microcirculation. *Nat Neurosci*, Aug;1(4), 286–289.
- Krishnan, V., & Nestler, E. J. (2010). Linking Molecules to Mood: New Insight Into the Biology of Depression. *American Journal of Psychiatry*, 167(11), 1305–1320.
- Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50(3), 249–258.

- London, M., & Hausser, M. (2005). Dendritic computation. *Annu. Rev. Neurosci.*, 28, 503–532.
- Lycan, W. G. (1981). Form, Function, Feel. *Journal of Philosophy*, 78(Jan), 24–50.
- Ma, Z., Stork, T., Bergles, D., et al. (2016). Neuromodulators signal through astrocytes to alter neural circuit activity and behaviour. *Nature*, 539, 428–432.
- Magistretti, P. J., & Allaman, I. (2018). Lactate in the brain: from metabolic end-product to signalling molecule. *Nature Reviews Neuroscience volume*, 19, 235–249.
- Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York: Henry Holt and Co.
- McIntyre, C. K., & Roozendaal, B. (2007). Ch 13: "Adrenal Stress Hormones and Enhanced Memory for Emotionally Arousing Experiences" in *Neural plasticity and memory: from genes to brain imaging*, pp 265–284.
- Mel, Bartlett W. (2007). 'Why have dendrites? A computational perspective', in Greg Stuart, Nelson Spruston, and Michael Häusser (eds), *Dendrites*, 2nd edn (Oxford, 2007; online edn, Oxford Academic, 22 Mar. 2012).
- Moore, C., & Cao, R. (2008). The hemo-neural hypothesis: on the role of blood flow in information processing. *J Neurophysiol.* 2008 May, 99(5), 2035–2047. Epub 2007 Oct 3.
- Nedergaard, M., Ransom, B., & Goldman, A. (2003). New roles for astrocytes: Redefining the functional architecture of the brain. *Trends in Neurosciences. Volume 26, Issue 10, Pages 523–530*.
- Ohta, T. (2001). Neutral Theory, in *Encyclopedia of Genetics*, commenting on King & Jukes (1969).
- Ott, T., & Nieder, A. (2019). Dopamine and Cognitive Control in Prefrontal Cortex. *Trends Cogn Sci.* 2019 Mar;23(3):213–234. Epub 2019 Jan 31.
- Panzeri, S., Macke, J.H., Gross, J., Kayser, C. (2015) Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences* 19(3) 162–172 S1364661315000030. <https://doi.org/10.1016/j.tics.2015.01.002>
- Perea G, Araque A. (2007) Astrocytes potentiate transmitter release at single hippocampal synapses. *Science.* 2007 Aug 24;317(5841):1083–6.
- Piccinini, G., & Maley, C. (2014). "The Metaphysics of Mind and the Multiple Sources of Multiple Realizability." In *New Waves in the Philosophy of Mind*, M. Sprevak, & J. Kallestrup (Eds.), 125–52. London: Palgrave Macmillan (2014).
- Pirritano, D., Plastino, M., Bosco, D., Gallelli, L., Siniscalchi, A., & De Sarro, G. (2014). Gambling disorder during dopamine replacement treatment in Parkinson's disease: a comprehensive review. *Biomed Res Int.* 2014:728038.
- Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences, Volume 10, Issue 2, February 2006, Pages 59–63*.
- Polger, T. (2004). *Natural Minds*. Cambridge MA: MIT Press.
- Polger, T. (2008). Two confusions concerning multiple realization. *Philosophy of Science* Vol. 75, No. 5, Proceedings of the 2006 Biennial Meeting of the Philosophy of Science Association Part II: Symposia Papers Edited by Cristina Bicchieri and Jason Alexander, pp. 537–547.
- Polger, T., & Shapiro, L. (2016). *The Multiple Realization Book*. Oxford University Press UK.
- Putnam, H. (1967). "The Nature of Mental States" (originally published as "Psychological Predicates"). In W. H. Captain, & D. D. Merrill (Eds.), *Art, Mind and Religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.
- Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.
- Pylyshyn, Z. (1980). Reply to Searle in Searle, J., 1980, 'Minds, Brains and Programs', *Behavioral and Brain Sciences*, 3: 417–57.
- Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., & Tenenbaum, J. (2021). Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* Nov 118 (45) e2105646118.
- Schultz W, Dayan P, Montague RR. (1997) A neural substrate of prediction and reward. *Science.*;275:1593–1599.
- Searle, J. (1990). Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association*, Vol. 64, No. 3, pp. 21–37.
- Shapiro, L. (2000). Multiple Realizations. *Journal of Philosophy*, 97(12), 635–654.
- Shapiro, L. (2008). How to Test for Multiple Realization. *Philosophy of Science*, 75(5), 514–525.
- Shapiro, L. A. (2004). *The Mind Incarnate*. MIT Press.
- Sharma, J., Angelucci, A., & Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature*, 404, 841–847.

- Sprevak, M. (2018). Triviality arguments about computational implementation, in Routledge Handbook of the Computational Mind (edited by Mark Sprevak & Matteo Colombo), Routledge: London, pp. 175–191.
- Steinert, J. R., Robinson, S. W., Tong, H., Hausteine, M. D., Kopp-Scheinpflug, C., & Forsythe, I. D. (2011). Nitric oxide is an activity-dependent regulator of target neuron intrinsic excitability. *Neuron*, 71(2), 291–305.
- Steinert, J. R., Kopp-Scheinpflug, C., Baker, C., et al. (2008). Nitric oxide is a volume transmitter regulating postsynaptic excitability at a glutamatergic synapse. *Neuron*, 60(4), 642–656.
- Stellwagen, D., & Malenka, R. C. (2006). Synaptic scaling mediated by glial TNF- α . *Nature*, 440, 1054–1059.
- Wei, Z., Inagaki, H., Li, N., Svoboda, K., Druckmann, S. (2019) An orderly single-trial organization of population dynamics in premotor cortex predicts behavioral variability. *Nature Communications* 10(1) 216. <https://doi.org/10.1038/s41467-018-08141-6>
- Turrigiano Gina. (2007). Homeostatic signaling: the positive side of negative feedback. *Current Opinion Neurobiology*, 17(3), 318–324.
- Wimsatt, W. C. (1986) *Developmental Constraints, Generative Entrenchment, and the Innate-Acquired Distinction*. pp.185–208 in W. Bechtel (Ed.), *Integrating Scientific Disciplines, Science and Philosophy*, vol 2. Springer, Dordrecht.
- Wimsatt, W. C. (2006). Reductionism and its heuristics: Making methodological reductionism honest. *Synthese Volume*, 151(3), 445–475.
- Wimsatt, W. C. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge MA: Harvard University Press.
- Yamins Daniel, L. K., & DiCarlo James, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci*, 19(3), 356–365.
- Zhang, J. M., Wang, H. K., Ye, C. Q., Ge, W. P., Chen, Y. R., Jiang, Z. L. & Duan, S. M. (2003). ATP Released by Astrocytes Mediates Glutamatergic Activity-Dependent Heterosynaptic Suppression, *Neuron*, Volume 40, Issue 5, Pages 971–982.

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