# *Nervous Systems, Functionalism, and Artificial Minds*

**Peter Godfrey-Smith**

University of Sydney

A talk given online to the NYU Mind, Ethics, and Policy Program, October 2023

## *1. Introduction*

I'll present a picture on some current topics: brains, computers, consciousness, animals far from us, the role of biology, and so on.[1] This picture is a bit different from familiar ones. I won't be able to argue for all of it here, though I will give some arguments. In particular, I will try to show that some arguments that have been used to support other views are not good arguments. I can at least motivate a rethinking. With that done, I'll offer a positive picture. Along with the philosophy of mind, I'll briefly make contact with ethical topics from time to time.

Here is a familiar view. Physical systems can have mental properties, including felt experience or consciousness, in virtue of their functional organization. The mental is "substrate-neutral" (or "substrate independent"). That is, the hardware does not matter, except in practical terms. What matters is the system's organization. You could realize the organization of a human mind in a computer system that does not contain proteins, DNA, and so on. Any substrate is OK if it can be suitably organized, and computers, through their programming, can end up organized in all sorts of ways.

Why believe this? Abstract arguments for some of the original forms of functionalism were supposed to tell us that causal roles are all that matter, and these might imply a kind of substrate-neutrality. There's also the great intuitive force of neural

[1] This text is fairly close to the talk given to the NYU Mind, Ethics, and Policy Program, and it retains the informal character of the talk. Additions are made in the footnotes, often in response to questions raised at the talk. References are very incomplete. Thanks to Jeff Sebo and Sofia Fogel for setting up and running the event.

replacement arguments, perhaps first given by Pylyshyn, and developed in detail by Chalmers.[2] Imagine a cell by cell replacement of your brain with artificial control devices that have a different make-up but play the same roles. As the replacement is done, your behavior does not change. You chatter away as usual. Does your experience fade as this happens? That seems at least an odd thing to be committed to. The organization is still in place, and perhaps that's all that matters.

Continuing within the familiar view, we can turn to some problem cases: invertebrate animals, computer systems running AI software, and robots. These are all uncertain in practical terms, as we don't yet know which features of our brains are the functionally important ones – we don't know exactly what has to be carried over into, or found in, those other systems. If we did a *perfect* simulation of a human brain, in every detail, in a computer system, then we'd know we had what is needed, but in practical contexts where less of "us" is in place, it's hard to know what's essential.

On this view, questions about animals far from us and questions about AI systems are pretty similar. They come down to the need to find the functional properties that matter to consciousness, as the physical hardware differences are not important in principle.

All this matters to ethical discussions. Many people accept a form of *sentientism*: being sentient is necessary and sufficient for moral considerability. (I am going to treat consciousness and sentience as the same thing here.) As we learn which functional properties matter for sentience, we learn which living things, and also non-living things, are – or will be, once they are built – sentient. Maybe some present-day AI systems are not too far away from a basic kind of sentience; perhaps the level of uncertainty is around the same place as it is with earthworms or flies.

That is the end of my sketch of the familiar view. I oppose quite a lot of this. I think that nervous systems are special, and biological properties probably matter to consciousness. There's more of a divide between the cases of distant animals and computers than the familiar view allows. The biological properties that I think are important are found in a lot of animals – I will look later at flies. For an artificial system to be a fly-like candidate for consciousness, it would need a different hardware from

[2] In *The Conscious Mind.* For Pylyshyn, see "The Causal Power of Machines," *BBS* 1980.

contemporary computer systems. It might not have to be alive, and perhaps need not be made of the same materials as us and flies, but it would have to be closer. The ethical side is affected: there's more reason to be ethically cautious about many far-from-us animals than about present-day computer systems, no matter what program they are running.

How can I make a case for this view in a short talk? I'll start by opposing the familiar motivations and arguments for substate neutrality, and then make a positive case for the importance of some particular "biological" properties of nervous systems.

## *2. Substrate neutrality and functionalism*

I think that philosophers (not only us, but us) have gotten used to talking about functional properties and hardware in a way that does not entirely make sense. This begins as an in-principle point, though later I will link it to empirical work. I want to credit Rosa Cao all through here. I was teaching a standard philosophy of mind course at Harvard around 2006, including "the long march," as Susanna Siegel called it – from dualism through behaviorism and the identity theory to functionalism. Rosa was sitting in as a postdoc, and at a certain point she started pushing back on a lot of things, especially around the alleged irrelevance of the biological details and hardware. Eventually I started to agree with her.[3] Today I'll be presenting my version of these ideas, but her influence was important.

Philosophers have gotten used to talking about functional properties and functional profiles in a way that supposes there is such a thing as a *perfect functional duplicate* of a person's brain. This is a physical system with *all* the "functional" properties of a brain, one that *does the same thing*. It is a perfect simulation or realization, in different hardware. I don't think this really makes sense. The functional similarity of two systems is a matter of degree. What a system, or a part of it, *does* can be understood in coarse or finer-grained ways.

The simplest way to make this point is with details of timing. Some facts about timing in a brain are certainly "functionally relevant" – both the time taken for things to happen as a whole, also the simultaneity of parallel processes, and so on. Some other

[3] Her main paper is "Multiple realizability and the spirit of functionalism," *Synthese* 2022, and we have a not-quite-done joint paper.

details of timing look like they probably don't matter much, but they're still *there*; the brain processes are still different if they are changed. Time tends to be ignored in classical functionalist discussions – there's the machine table, or imagined network, and if it gets things done, in some amount of time, then it gets them done. (Microsoft Word runs faster on this machine than that one, but it does run on both.)

How much fidelity of timing does there have to be in a "perfect" duplicate? That is an ill-posed question. Suppose something takes a millisecond longer, or the synchrony of two parallel processes is not quite the same.... What there is, in these cases, is functional *similarity,* to various degrees.

Timing can make the point, and so can other details – a few more ions here or there. Also relevant is whether two actions are "the same" across two occurrences or two systems. That is also a matter of degree – you lift your arm a millimeter higher this time than you did last time.... All sorts of details can differ, and often the details don't matter much, but they are still there.

Chalmers in his *Reality+* book talks about perfect duplicates – "If there's a physical process in the brain that *makes a difference* in how the brain functions, it will be simulated" (p. 287) – but this raises the same question. Is another millisecond here or there a "difference"? Yes, it's a difference, but a tiny one. It's a difference in experience, too; if some process that is the basis of experience goes on for another millisecond in this other system, then that is a difference.[4]

---

[4] At the NYU talk, Chalmers raised a passage from *The Conscious Mind* (p. 331) where he claims, in relation to replacement scenarios, that "when it comes to duplicating our cognitive capacities, a close approximation is as good as the real thing." His argument is that in biological systems, random "noise" processes play a role (greater than the role of any analogous processes in a computer). When the biological system performs some operation, the outcome is never entirely reliable and will instead fall within a band of possibilities. An artificial duplicate of the biological system only has to give a result somewhere in that band. The duplicate's output might depart from what the biological system actually does, on some occasion, but the biological system could just as well have produced the same output as the duplicate, if noise had played a different role. When a duplicate gives a result within the band, it is doing "as well as the system itself can reliably do."

In response, it is true that this role for noise is an important micro-functional feature of living systems. In addition, neurons change what they do as a result of their normal operation, they don't respond to the "same" stimulus twice in the same way (see "Mind, Matter, and Metabolism" for references). The "rules" or the "program" being followed are always changing as a result of the activity of the system itself and its embedding in other biological processes. Over time, the effects

This bears on neural replacement arguments. Your brain cells might be slowly replaced by physically different units that "do the same thing," we're told. But they won't do exactly the same thing. As the replacements are done, small differences will accumulate. Some people would deny this – they would say it's possible to have no differences when the physical substrate is changed. Do they mean really *no* differences? Not millisecond-scale differences? It is surely more likely that as you replace neurons, the system changes micro-functionally, and it will also change micro-behaviorally – in timing, fidelity of repetition, in all sorts of details. What is going on, both inside and outside, is different.[5] This means there's no reason to believe that experience is unaffected in a slow replacement. First a little, then a lot. This could include all sorts of transformation and fading.

In a moment, I will compare my view to Ned Block's. He also rejects replacement arguments of this kind (see his reply to Tye in the *Blockheads* collection). Block sees these arguments as question-begging. Even if a hardware replacement made no differences to behavior at all, it is question-begging to say this shows that felt experience cannot change.[6] Maybe it does change – that is the whole question.

I more-or-less endorse Block's point. But under the mistaken terms of the original arguments, replacement scenarios did put a lot of pressure on more biological views. We

---

of these factors will accumulate and compound – a comparison of what a living system and a duplicate might do in a single operation doesn't capture their importance. I see all this not as a "lowering of the bar" that enables us to keep talking in a rough way about functional identity, but another functional difference between living and artificial systems.

[5] In some neural replacement scenarios, the replacement is a slow irreversible process. In others ("dancing qualia" in Chalmers), there's an ongoing switching back-and-forth between biological and artificial controllers. The latter is easiest to think about here. The argument is that the hardware switching leaves the functional organization "the same," and I am denying this for fine-grained functional properties. Whatever it means to say you would not notice "from the inside," differences will be visible from the outside.

In the case of a slow irreversible process, the passage of time and accumulation of experience will have effects on the functional profile of the system whether there's a replacement going on or not, so talk of behavior and cognition being "unaffected" or "unchanged" has to be seen as comparing a history in which the biological hardware remains in place with a history in which the replacement is done, and saying that there's no difference between them. In this case, I am saying that there are differences.

[6] See note 5 again in relation to the idea of "change" here.

might decide that it's resistible pressure, but it's a lot: the system, we're admitting, works in *exactly the same way* despite the physical changes. I think the point about the nature of functional properties, their grain-dependence, is more basic as an objection.[7]

We're freed-up from thinking that substrate-neutrality is "compulsory" – something we have to accept and work with. The usual arguments for it are not good. But that does not mean there's not a lot of truth in the view. It does not mean that that you could not reproduce *everything that matters much* in a physically different system, such as a computer of present-day design. It doesn't show that biological details are actually important.

That's right; it frees us up to ask those questions again. Might the biological details matter? All I can do on that point, in this short discussion, is present a picture. Here is my view (as far as I've got with one), of the biological and physical basis of subjective experience.

### *3. Biology and consciousness*

My view has two parts with uncertain relations between them. The problem of consciousness is sometimes expressed (for example by Nagel) using the idea of point of view, or subjectivity. I embrace this. Animal evolution builds systems with points of view. It does so by building sensing, action, and certain kinds of processing that links them. All this takes us some distance with the problem. The other part of my view, more relevant here, is some claims about nervous systems. Nervous systems combine two

[7] Some quotes from the Block-Tye exchange in the *Blockheads* book. Chalmers quoted and endorsed by Tye: "As long as the chip has the right input/output function, the replacement will *make no difference* to the functional organization of the system." Tye: "there seems no reason why the input-output functions could not be duplicated using silicon chips. The question to be addressed is whether the phenomenology would change under the above scenario." Block: "There are many mechanisms of neural information transfer that on the face of it may be difficult or impossible to simulate in real time in a small space. Neurons affect other neurons in part by many types of complex mechanisms (for example, slow profusion of neurotransmitters into extracellular fluid). And some transfers of information work via direct connections between neurons ("gap junctions") through which many types of molecules can flow from one neuron to another-rather than via a synapse. But I put these issues aside for the moment and assume that the scenario that Chalmers describes is indeed possible."

features. First, there are cell-to-cell interactions mediated by synaptic connections – the network properties of nervous systems. Second, there are what I will call *large-scale dynamic properties*, such as the oscillations picked up in an EEG (brain waves). These are partly distinct from "spikes, or action potentials, as they based more on ion movements across membranes that are not strong enough to initiate a spike. In a tradition at least 30 years old (perhaps older), some writers have claimed that these large-scale dynamic properties play a role in experience-relevant, or at least subjectivity-relevant, aspects of cognition.[8]

Synchronized brain oscillations appear to play an integrative role in sensory experience, a role that has been studied especially in case of vision. They also have roles in selective attention, sleep/wake cycles, and anesthesia.[9] The apparent role of these features in experience-relevant phenomena is not only seen in us; it's also found in animals far from us. I've been influenced by Bruno van Swinderen's work, in particular. His lab works on flies (*Drosophila*). I'll sketch two of their experiments. An earlier round of work showed an association between oscillatory activity in the beta range (20-30 Hz), and what looks like *selective attention* on objects in flies.[10] The experimenters moved a shape through the fly's visual field and noted both a beta-range response and a slower-wave one. The former seemed to be associated with attention to objects. For example, the beta-range response was modulated by reward (unlike the slower wave). It was sensitive to shape, but not figure-ground illumination properties (and the slower wave had the opposite combination). In a sleep-like state, the beta response was much reduced while the slow was not.

A newer paper (with Martyna Grabowska as first author) showed the flies visual stimuli that had distinct flicker rates, along with other differences, such as size, which flies are known to care about.[11] Those flicker rate differences can be seen in the brain, when the fly is attending to an object with a particular flicker rate. These flicker "tags"

---

[8] (References to add to Singer, Crick and Koch, etc.)

[9] (References to come – Melloni, Singer 2018, van Swinderen, others.)

[10] See van Swinderen & Greenspan, "Salience Modulates 20–30 Hz Brain Activity in Drosophila," 2003.

[11] Grabowska et al., "Oscillations in the Central Brain of Drosophila are Phase Locked to Attended Visual Features," PNAS, 2020.

are not the oscillations being studied – they are slower. But the flicker tags can be correlated with the faster, potentially important ones. In one experiment, two objects, one large and one smaller, were presented in the fly's visual field. The files usually prefer larger objects, and in the control condition, their brains were more matched to the large object's flicker rate. Optogenetic methods were then used to activate a reward circuit in the fly's brain, and associate reward with the smaller object. Then, when both objects were in the visual field, the fly was more likely to attend to the smaller one. Stimulation of the reward circuit made the smaller one more salient, more interesting.

What relation does this have to beta oscillations? Endogenous beta oscillations can be phase locked (linked in their timing) to one object tag frequency or another. If you make the smaller object more interesting, with the reward circuit, its flicker tag gets synchronized with the endogenous beta oscillations.

This is an example. There is work on active sleep and anesthesia in animals like this, as well.[12] All of it suggests an experience-relevant role for large scale dynamic patterns. Nervous system activity combines point-to-point, network interactions with more diffuse, somewhat holistic electrical phenomena. This combination, I conjecture, is pivotal to the biology of felt experience.

I've presented just one angle on one kind of data, but *if* we were to read a message off this, we might do it as follows. Architecturally, there are lots of options for animal brains – cortex, no cortex.... Features like that don't matter to the existence of some experience-relevant properties: attention, sleep/wake distinctions, and so on. And when we look at nervous systems, some things we find that *do* have apparent connections to these experience-relevant properties are very "biological." They involve the combination of features in neural activity that I mentioned earlier. The general biology of nervous systems makes these features possible, and there's a lot of flexibility on the architectural side. Animals with different bodies and histories handle the architectural side in their own ways, while conserved features of nervous system activity itself seem to be doing experience-relevant things in many of them.

[12] See, for example, Van De Poll & van Swinderen, "Balancing Prediction and Surprise: A Role for Active Sleep at the Dawn of Consciousness?"

Next I will make a more phenomenological point. I want to suggest a mapping between this view of nervous system activity and a feature of experience itself. I suggest that what is basic to ordinary human experience is what I call *experiential profiles.* An experiential profile is a total way things feel to someone at a moment. These profiles are inherently multifaceted, gestalt-like. There's something in attention, while other things are in the background. There's bodily awareness, and the whole is modulated by things like mood and energy level – usually far from attention, but part of *what it's like* for the subject of experience at that moment.

From here, I look back to that picture of nervous system activity. When neural activity has the form of large-scale dynamic patterns that are modulated, a range of different senses, and other factors, will naturally tend to affect the present state.[13] An internal state strongly affected by what the animal is seeing will also be affected by tactile sensing, proprioception, and more. The gestalt-like character of experiential profiles, with the ebb and flow of salient modalities and attention, also the role of mood and energy level, and so on, is naturally explained by this view of nervous system activity. A particular combination of nervous system features has something of a bridging role with respect to the explanation of consciousness.

I'll call the package of ideas the NDS view, for *neural dynamics of subjectivity.* This is a "biological" view of consciousness, in part. Block has also defended such an approach, in brief sketches.[14] A biological feature he suggests might matter to consciousness is the continual transitioning in nervous systems between electrical and chemical information processing. A back and forth between these is a distinctive feature of nervous systems.

Here are a few thoughts about this idea. The property Block emphasizes might be a near-inevitable way of doing things once you get beyond very simple nervous systems. As Gáspár Jékely has discussed, a "chemical brain" that uses the broadcast of many signaling chemicals, and lacks targeted projections between neurons, could work well in some ways, but perhaps only in simpler nervous systems and where speed does not

[13] A number of people have expressed versions of this idea, from Mac Passano in 1963. (References to come.)

[14] See his "Comparing the Major Theories of Consciousness," and the *Blockheads* replies.

matter much. On the other side, a purely electrical system, with physical connections between cells and "gap junctions" connecting them, one lacking chemically mediated synaptic connections, might be very fast, but it might be hard to achieve useful plasticity in such a system.[15] It might be difficult for such a system to re-wire itself through experience. The modification of chemical synapses is a good way to achieve experience-dependent plasticity.

If an electro-chemical back and forth is near-inevitable, that is not a problem for Block – this feature has to have a evolutionary story of some kind behind it. But what I don't yet see in this proposal is why the back-and-forth is a gap-bridging sort of property, in the sense of Levine's "explanatory gap" – how it links to the physical explanation of experience itself, rather than just being something in brains that makes general sense from a design point of view. Some of this gap-bridging is what I tried to show for my view, above, in that discussion of a link between the perturbation of global states of activity and the multi-faceted nature of experiential profiles.

Block, on the other hand, has suggested to me that if large-scale dynamic properties of neural activity matter a lot to consciousness, this should be more obvious than it seems to be. Our brains are continually being exposed to flickers at different frequencies, that might affect oscillatory patterns, in modern electric lights, and the like. Why doesn't this aspect of our environments have obvious effects on experience? Good question.

### *4. Discussion*

I argued first that we need not believe in substrate-neutrality. Consciousness might depend on the biological side, on the hardware. But does it? There's some reason to think so. Nervous system activity has a combination of features that is both different from what is seen in artificial systems (as far as I know), and also experience-relevant. These

[15] See Jékely,"The chemical brain hypothesis for the origin of nervous systems," 2021, and, for another side of the coin, Burkhardt et al., "Syncytial nerve net in a ctenophore adds insights on the evolution of nervous systems," 2023.

features are not easily exported into a different machine. Suppose all that is true. What follows?

First, artificial systems need to have something *physically like* the features of nervous system activity that I've been discussing. The argument is not that no artificial system could be conscious, but such a system would have to be set up a fairly brain-like way. It might indeed be possible to build something with analogues of the relevant features; this would be a system with large-scale, diffuse patterns of activity akin to those seen in brains, along with networks of interaction and ways for the system's state to be perturbed by various stimuli. The NDS account also gives a role to more traditional schematic or "functional" properties involving perspective and point of view. These properties need to be *realized* in the artificial system, not just modeled. Such a system may well be something we could build, but ordinary computers doing smart things, like the large language models, are not like this. They lack the schematic perspective-related properties as well as the brain dynamics. Robots with good senses are closer to achieving the perspective-related properties, and in that case the issue is the physical nature of the control system.

What about animals far from us – insects, jellyfish, earthworms? The issues they raise are very different from the ones that arise with AI and computers. These animals have nervous systems, and have the perspective-related properties as well. They have this in much smaller nervous systems than are seen in mammals like us. But the path that led us here ran partly through work on flies. Flies have large-scale dynamic properties of an experience-relevant kind in brains on a scale of 200,000 neurons or so. That is smaller than a bee and much smaller than an octopus (500 million neurons total, with about a third of those in the brain). Even jellyfish-like animals, including *Hydra*, have these neural patterns. *Hydra* has at most a few thousand neurons. I don't know about nematodes.

In those animal cases, we confront empirical uncertainties, functional differences at various scales, and the role of simplicity versus complexity. In relation especially to that last point, we are likely to end up with a gradualist view of consciousness, one that does not include a sharp line between yes and no. That is suggested by the evolutionary side. Gradual origins for consciousness are likely, and that suggests, though it does not

imply, a graded presence of consciousness across different animals now.[16] Gradualism complicates the situation, both for animals and for artificial systems, too.

My conclusions can *roughly* be expressed in simple terms: Nervous systems are special. Plants and bacteria are out. The way that nervous systems are special brings a lot of animals *in*, including neurally simpler ones very far from us. AI systems need new hardware. Cerebral organoids are a different matter, and these may pose practical problems for us quite soon, as may hybrid neural-artificial systems, when they have enough on the neural side.

That is the rough version. But the fact that a sharp distinction between conscious and non-conscious systems is unlikely makes all of this more complicated. This shows up also on the ethical side. If we assume sentientism, we can read off some initial conclusions from what I have above – about plants, at least some invertebrates, present AI systems, and so on. But in the case of neurally simpler animals, perhaps including earthworms and many insects, we have to be ready for a gray area between conscious and non-conscious (sentient and non-sentient). If sentience is graded, what becomes of sentient*ism*? The situation would still be pretty simple if there was a sharp line taking a system to a *minimal yes*, and various gradations from there (I call this weak gradualism). Then we could say: given the minimal yes, you are considerable, and the way in which you are is affected by a richness or complexity gradient – or more likely, several such gradients. But the view I think likely is one without a sharp line dividing non-conscious cases from a minimal yes (this is strong gradualism). If sentience has a graded presence in this way, what happens to moral consideration?

---

[16] See "Gradualism and the Evolution of Experience," *Philosophical Topics*, 2020. There's more on most of the themes of these last few paragraphs in my 2023 Whitehead Lectures. https://metazoan.net/108-whitehead-lectures/