What Difference Might Biology Make?

Peter Godfrey-Smith
University of Sydney

A talk given at the "Distribution of Consciousness" Workshop, Galápagos Islands, Ecuador, 2026. This text is close to the talk as it was given, except where noted. I've added footnotes with extra thoughts, comments about discussion after the talk, and so on. There's also an Appendix about AI. Many thanks to the International Center for Consciousness Studies (ICCS) for organizing the floating workshop. The main text is about 4400 words.

1. Introduction

I'll look at some candidate "biological" properties that might matter to consciousness. I think these are real candidates, and they also have an in-principle role: they give us a sense of *how* biological and hardware-related features might matter.

The biological properties I'll discuss are probably not impossible to have present, in a decent form, in a computational artifact. So even if I am right about the candidates, they don't push toward the view that "nothing non-living could ever be conscious."[1] But, I take it, these features are not in AI systems now, are not seen as priorities, and would not be easy to introduce. It would be a new hardware project.

The topic is fairly well set up by recent discussions of "biological naturalism" and "computational functionalism."[2] Ned Block:

> Computational functionalism claims that executing certain computations is sufficient for consciousness, regardless of the physical mechanisms implementing those computations. This view neglects a compelling alternative: that subcomputational biological mechanisms, which realize computational processes, are necessary for consciousness.[3]

[1] Anil Seth: biological naturalism is the view that "consciousness is a property of only (but not necessarily all) living systems" ("Conscious artificial intelligence and biological naturalism").
[2] "*Computational functionalism:* Implementing computations of a certain kind is necessary and sufficient for consciousness, so it is possible in principle for non-organic artificial systems to be conscious" (Butlin et al., "Consciousness in Artificial Intelligence...").
[3] Block, "Can only meat machines be conscious?" TRiCS. Emphasis added.

That is not exactly how I see things, though it's close. The distinction between computational and *sub*computational might be what matters, but maybe not. Some important biological processes might be *sort-of computational* – quasi-computational – and far from what actual computers contain. And all this might be a matter of "levels," but might not be.

Part of the question here is an old one of multiple realizability (or "substate neutrality"). How hard is to make what's essential to minds in a different material? But two extra concepts are the table: computation, and the biological. In the case of computation, familiar analyses make use of the ideas of symbol manipulation, algorithmic processes, and the mechanization of logical relationships. I accept a version of this view: There are paradigm cases of computation where those features apply, accompanied by a cloud of non-paradigm cases, and there's also some interpreter-dependence. Is the regulation of gene expression in (all) cells a computational process? *Sort-of yes*, as it has Boolean features (it is full of *ands*, *ors*, and *nots*), and has these featues by design.[4] This was the Earth's first computational process. But it's relatively simple, approximate in its match to logical operations, and not *just* computational – there's more going on. (When we can call a process "computational," we feel entitled to throw all kinds of detail away.)[5]

What is the biological? Here, a person might mean

(1) *comprised of, or materially similar to, biological structures* in actual living organisms – with proteins, membranes, etc.

Or they might mean:

(2) *having the general properties that make something alive* (where for me, this bifurcates into metabolic and reproduction-related features).[6]

---

[4] *Genes and Signals*, Ptashne and Gann, 2002. This is a case that's relatively simple but not, I think, much affected by interpreter-dependence. On that side, it's about as straightforward as it could be: Boolean operations are installed by evolutionary design.

I think of interpreter-dependence as a general feature of the situation, not something that only applies outside the paradigm cases. But there's an interaction there; sometimes a computational interpretation is *so* well motivated that this feature recedes, and sometimes it's not.

[5] This sentence is an addition, as the idea became vivid during discussion. The language of computation induces us to look to abstract structural features and ignore others.

[6] *Living On Earth*, chapter 2.

I'll talk about these as the biological in senses 1 and 2.[7] Block's proposal about the possible importance of a combination of chemical and electrical influence in nervous systems is a biological-1 proposal. Anil Seth defends "biological naturalism" in a different way. He makes use of the "predictive processing" framework of Friston, Andy Clark and others, which holds that the function of perception, cognition, and even action is to minimize surprising or disruptive experiences.[8] For Seth, this tracking and control of surprise is a feature of life in general. It is tied to the self-preserving and self-defining aspect of life. This is biological-2; more abstract, not tied to the materials.

My candidate is biological in sense 1. I don't deny that some sense-2 properties might be important as well.[9] (Here is an acronym collection for this written text: NS = nervous system; EP = experiential profile; LSD = large-scale dynamic.)

---

[7] David Chalmers, in comments on this talk at the workshop, compared this 1-2 distinction to the distinction between "role" and "realizer" functionalism in the philosophy of mind. I see the comparison, but don't think it's very close. I take it that role and realizer functionalists don't disagree very much. Janet Levin in the SEP: "if in humans this functional role [of pain] is played by C-fiber stimulation, then, according to this functionalist theory, humans can be in pain simply by undergoing C-fiber stimulation. But there is a further question to be answered, namely, what is the property of pain itself? Is it the higher-level relational property of being in some state or other that plays the “pain role”in the theory, or the C-fiber stimulation that actually plays this role?" The choice can have downstream consequences. In the case of the "biological," though, the candidates for a biological$_1$ property that might matter to consciousness, at least so far, are not realizers of roles essential to being alive. They are contingent features of the inherited hardware of animals, especially within nervous systems. They need not have analogues in other forms of life, actual or possible.

[8] In an in-press commentary on Seth's BBS paper, I criticize the predictive processing framework in its strong forms. https://petergodfreysmith.com/wp-content/uploads/2026/08/PGS-Commentary-on-Seth-E2-website-preprint.pdf. The main idea (from the commentary): Effective action can reduce uncertainty in some cases, but can also embrace it, and the sensory feedback accompanying action is sometimes of minor importance. Action does not serve prediction and perception; the functional relationship runs in the other direction.

[9] Some people think that living activity is associated with the *reality of value* for a being, in a way relevant to whether that being can have experiences. This idea is part of Seth's picture – living systems have "skin in the game." It's not tied to the predictive processing framework, though, and might be important outside it. That, also, is a biological$_2$ candidate.

2. Nervous system activity

What is nervous system (NS) activity? Here is a standard view: although there's a great deal going on in nervous systems, the core of it for understanding the mind is a collection of interactions within a network – with nodes, targeted connections between nodes, excitation and inhibition of action potentials (spikes), and modification of synapses. These features are also taken to be easy to realize in a computer (though I'll come back to this issue).

Here is an alternative view. NS activity has two sides (at least). There are networked, point-to-point influences, and more diffuse electrical activity that is often rhythmic. The most conspicuous aspect of this second side of NS activity is seen in an EEG, but this is the tip of an iceberg.

I'll use the term "large scale dynamic" properties (LSD activity, or just LSDs) for this second side. LSD activity interacts with network activity. Each affects the other. But LSDs (or at least these ones) are not reducible to patterns in spikes – they are not the sum of spikes.[10] In the case of oscillations (again, perhaps not all there is), their basis is subtheshhold ion flows across cell membranes.[11] They partially depolarize or hyperpolarize neurons, often in a coordinated way.[12]

These patterns are conspicuous in a human EEG because parallel cortical cells near the skull create a strong signal. But similar LSD patterns are also seen in flies, bees, octopuses, snails... even *Hydra* (a tiny jellyfish-like animal). They are all over the

---

[10] György Buzsáki's 2006 *Rhythms of the Brain*: "the contribution of action potentials to the local field and especially to the scalp EEG is negligible."

> [W]e distinguish excitatory postsynaptic potentials (EPSPs; or currents, EPSCs) from inhibitory postsynaptic potentials (IPSPs; or currents, IPSCs). Compared to the fast action potentials, membrane potential changes associated with EPSPs and IPSPs are several-fold smaller in amplitude. However, they last for tens of milliseconds. This latter property is critical for understanding the generation of EEG activity.

Destexhe and Sejnowski: "EEG activity is not generated by action potentials" ("Interactions Between Membrane Conductances..." 2003).
See also, on my blog: https://metazoan.net/133-rhythms-2/

[11] Glia are part of the picture here and I've not yet looked at them.

[12] The term "ephaptic coupling" is used in a few ways (it seems to me), but generally refers to influence of one neuron on another through changes in the local electric field, as opposed to a synaptic connection.

animal kingdom. In more distant animals (without skulls) they are studied with invasive LFP (local field potential) methods.

LSD patterns are *associated*, at least, with *consciousness-related distinctions* (sleep/wake, anesthesia, attentive/passive). The rhythms occur at different rates, faster and slower. The slowest waves (delta) are prominent during quiet sleep. Much faster oscillations (gamma) are visible during focused attention. Others are in between (alpha, beta).

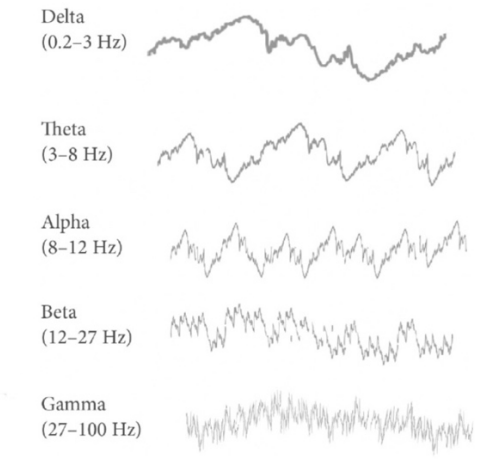

FIGURE 19: ***A depiction of the primary types of brainwaves, categorized according to frequency of oscillation.***

From Fields, *Electric Brain*

These associations with consciousness-related phenomena also hold in other animals, including the flies. An apparent role for beta oscillations in attention (or at least a fine-grained association) is conserved across animals (us and flies) whose last common ancestor was a flatworm-like animal living before the Cambrian. In octopuses, a distinction between quiet and active sleep, as in us, is reflected in oscillatory differences.

3. What do LSD processes do?

Here are some options (not all of them exclusive).

Option 1. Nothing. They are an epiphenomenon, a byproduct, the hum of the machine, mere "fumes."[13] This has been a common view. It's consistent with those associations with sleep/wake distinctions, etc.[14]

Option 2. General support (of a kind correlated with sleep/wake patterns, etc.)

Option 3. Organization of activity – timing and synchronization

LSD patterns may help to organize the more familiar interactions between neurons (excitation, spikes), especially in relation to timing. Here is Wolf Singer:[15]

> The purpose of this brief review was to highlight the importance that precise timing relations among the discharges of distributed neurons may have in the encoding of semantic relations, the dynamic gating of transmission and the formation of associations by learning mechanisms. Oscillations and synchrony... are ideally suited to endow neuronal responses with the precise temporal structure required for the realization of these functions.

A person might say: this could be very practically important, like a clock in a computer, but it's *ancillary*, a platform for the cognitively fundamental activity (spikes, network influences). Perhaps, and I'll come back to this point. But "like a clock" is also somewhat misleading at the outset. LSDs are not like an app that gets added on. This oscillation seems to be an old and deep feature of nervous systems, due in part to the features of excitable cells themselves, and also perhaps to inherent tendencies that oscillators have to align their activities.[16] (Reply: "I don't care; you could still do it another way" – again, I'll come back to this.)

---

[13] A debate entitled "Gamma – Fumes or fundamental?" took places at the Society for Neuroscience in 2018 (see Fields, *Electric Brain*).

[14] Buzsaki emphasizes conservation of oscillatory patterns and rates over very different size in mammals. Not likely epiphomenal, he says, given the size differences. How much more so when we include invertebrates.

[15] "Neuronal oscillations: unavoidable and useful?" 2018.

[16] There are two things to explain: why subthreshold oscillations happen in cells at all, and why they coordinate on a large scale (see a note for some details.) On the first, neurons themselves

An option at the other end of the scale:

Option 5. A medium of analog computation of their own
This is advocated especially by Earl Miller and his group at MIT: slower oscillations provide a frame (a "stencil") and faster ones are content-bearing signals.[17]

There's also an option "between" 3 and 5.

Option 4. Nervous system activity includes both aspects
This sounds bland, but it does has contrasts. LSDs are not ancillary, secondary, or merely supportive. NS activity has these two sides. Both are functionally important – important to what the system does. If NS activity is the basis of felt experience (if "sensations are brain processes") then *this is part of the basis.* It's not ancillary to the basis, but on the same footing. If one is a dualist and is committed to extra laws relating consciousness to brain activity, then in that weaker sense of "basis," LSDs can be part of the basis. (Again, I'm emphasizing oscillations here but there might be more as well.)

4. Consciousness

My aim is to develop that option 4 view, especially in relation to consciousness. Why might the LSD side of neural activity matter? The best known idea in this vicinity is the claim is about the possible role of 40 Hz oscillations in "perceptual binding" – in getting colors and shapes lined up properly in visual experience, for example (Singer,

---

can have an intrinsic oscillatory tendency (Llinás: "some central neurons have ionic conductances organized to endow them with electrical autorhythmicity. In addition, in many neurons the kinetics of these ionic voltage-dependent conductances are such that the cells may respond preferentially to inputs at a certain frequency or frequencies, that is, they behave as resonators") but it can also be a network product. On the second, there's an intrinsic tendency of oscillators to align with each other, something first noticed by Christiaan Huygens with clocks in the 17th century. Wolf Singer and Alison Hanson, among others, make this Huygens connection. See "Unavoidable..?"

[17] Some relevant Miller papers (all with collaborators): "Working memory control dynamics follow principles of spatial computing," "Cognition is an emergent property," Analog Cognition and Consciousness."

Melloni, Crick and Koch). I am not sure where that proposal has ended up, and I will look at others. Most of the empirical ideas here seem to share the same flavor.

Here is an example: Hipp et al. (2011), "Oscillatory Synchronization in Large-Scale Cortical Networks Predicts Perception"[18]: An ambiguous stimulus can be perceived in two ways. The stimulus is two identical bars approaching, meeting, and separating. This can be perceived as them "passing" through each other or "bouncing" off each other (and the subject also hears a click, which might encourage the latter). People tend to see either option, at roughly 50%.[19] A particular oscillatory pattern, when it's seen in the brain, predicts that you'll see it in the bounce way – which seems richer and more object-ish to me (the authors do not say that). The LSDs are tracked with EEG.

> when large-scale beta- band synchronization was enhanced between frontal, parietal, and extrastriate areas, subjects were more likely to perceive the same sensory stimulus as bouncing rather than passing.
>
> For identical sensory stimulation, the subjects' percept was predicted by intrinsic fluctuations of long-range [beta] synchronization directly preceding the sensory ambiguity.

There are lots of results linking beta or gamma synchrony to a perceptual phenomenon (finding a hidden object, resolving ambiguity). In the Hipp case there is a predictive relationship. The beta synchrony precedes the subject's report of what is experienced. On the other hand, you might say that this just reflects general intensity of attention, which tends to produce the 'bounce' interpretation? Maybe.

This is an EEG paper. Again, EEG does not record spikes (see footnote 10). I suspect people sometimes see papers like this and think "OK, the synchronization of spikes is important." It might be, and it might be enabled by the LSD patterns. Spikes

---

[18] *Oscillatory Synchronization in Large-Scale Cortical Networks Predicts Perception*
Hipp, Engel, and Siegel (2011), "Oscillatory Synchronization in Large-Scale Cortical Networks Predicts Perception." *Neuron* 69: 387–396.

[19] Details from the SI: "For [the] screening, subjects performed 100 audio-visual trials. For the main experiment we accepted subjects that had at least 30% 'bounce' and 30% 'pass' trials in the screening session (~60% of all screened subjects). This ensured a sufficient number of trials for both percepts for the data analysis. In the main experiment, each subject performed 500 audio-visual trials, followed by 100 visual-only control trials." I don't know what role the "control" trials played; I need to ask them.

are probably being *organized* by all this. But what the paper is tracking is not spikes. This is an "option 4" point: neural activity includes both sides.

In many papers, the tendency is to argue that coherence, unity, and the organization of cognition and experience have a relation to LSD activity. This can happen either in one modality or across several. In the Hipp et al. study, there is some role for multi-modal processing, the audible click. A second oscillatory synchronization (gamma frequency) was seen just in the left hemisphere in bounce-interpreted trials, and the authors interpret this as related to cross-modal integration. [20]

Multi-modal integration is where I want to go next. Here's the abstract of a review paper (Senkowski et al. 2008):[21]

> Picture yourself on a crowded sideway [sidewalk?] with people milling about. The acoustic and visual signals generated by the crowd provide you with complementary information about their locations and motion which needs to be integrated. It is not well understood how such inputs from different sensory channels are combined into unified perceptual states. Coherence of oscillatory neural signals might be an essential mechanism supporting multisensory perception. Evidence is now emerging which indicates that coupled oscillatory activity might serve to link neural signals across uni- and multisensory regions and to express the degree of crossmodal matching of stimulus-related information. These results argue for a new view on multisensory processing which considers the dynamic interplay of neural populations as a key to crossmodal integration.

---

[20] Hipp: "a left hemispheric network consisting of central and temporal regions that showed significantly stronger high gamma-band coherence (74–97 Hz) for bounce than for pass trials .... This perception related gamma-band synchronization started before and peaked around the time of bar overlap... The difference in coherence was caused by an increase during bounce trials... and a decrease during pass trials.. relative to the average prestimulus baseline. We confirmed the left lateralization of the network: Coherence did not differ between bounce and pass trials at the corresponding locations in the right hemisphere..." ...
"Consistent with this evidence, fluctuations of synchrony within the gamma network did not only reflect the subjects' percept of the ambiguous stimulus but also predicted interindividual differences in the cross-modal integration of auditory and visual information. Enhanced synchronization was specifically associated with the cross-modally more integrated bounce percept." [left side cf object-recognition]

[21] "Crossmodal binding through neural coherence: implications for multisensory processing." Daniel Senkowski et al, *Trends in Neurosciences*, 2008.

Perhaps this combination (LSDs and network interaction) is how the brain does what it does, and the LSD side is important in achieving certain kinds of integration. And perhaps it *feels* like something to have a brain doing this – doing this, and *doing it this way.* That's the idea I want to look at next.

The claim I want to develop is not that LSDs are *the* basis of consciousness, are a "magic bullet" in relation to the problem, or that qualia are electromagnetic clouds generated by LSDs. (Maybe they are, but that is not my claim.)[22] The idea, instead, is that nervous system activity as a whole, which is the basis of consciousness (at least in animals), includes two aspects (two at least). It includes network-based point-to-point influences and LSDs that tie things together. Network activity and LSDs also differ in how readily they can be realized in a computer.

My next step includes an idea that many will deny. What is ordinary experience like? In some views, especially the workspace tradition, it is essentially a highlighting of representations, a selective deployment and presentation of them. This works roughly one at a time, though sometimes with rapid switching between items.

Another tradition holds that ordinary experience has a multi-faceted character, with wholeness or unity also present. Remember the Senkoswki et al. quote:

> Picture yourself on a crowded sideway with people milling about. The acoustic and visual signals generated by the crowd provide you with complementary information about their locations and motion which needs to be integrated. It is not well understood how such inputs from different sensory channels are combined into unified perceptual states.

For Stan Dehaene, in particular, I take it that this passage requires a certain amount of reinterpretation, perhaps even push-back. We are cognitively making use of visual and auditory signals, and integrating them, yes, but in the conscious workspace "there can be only one."[23] Only one item can be in there at a time, or perhaps a number close to one (as I've heard some people say). And again, switching can be rapid.[24] Stan has

---

[22] This paragraph is added to this version, based on discussion after the talk. I need to emphasize all this, or a "magic bullet" impression arises.

[23] (*Highlander*, 1986). See Dehaene, *Consciousness and the Brain*. I discussed this feature of Dehaene's view in my LRB review of his book, “Do Squid Feel Pain?”

[24] How rapid can switching be? In his talk at the Galapagos workshop, Dehane said that the processing of an item in the workspace takes at least something like 150-200 milliseconds, so a

recently reasserted this feature of his view in his commentary with Naccache on the paper exploring whether the AI system Claude's "J-space" is similar to a workspace).[25] Some footnotes below have further comments on Dehaene's view (I was the commentator on his talk at the workshop). Here, I will move ahead within the other picture, the one in which experience tends to be multi-faceted at each moment.

Within that picture, here is a way to describe ordinary experience: at each moment we have an e*xperiential profile* (EP). Something will usually be highlighted by attention, and many other elements – or facets, a better term – will be in the background, but making some contribution to what it feels like to be you at that moment. These include background sounds, the lights of other cars approaching David Armstrong's distracted driver, energy level, mood.... There's a continual shifting of their prominence in the profile, but many are playing some role at each moment.

Breadth of experiential profile does not imply *naivete of presentation.* It remains true that you can fail to notice the man in the gorilla suit in the famous experiment, and he might not be part of your EP at all.

I suggest that the view of NS activity I'm working within provides a natural explainer for these features of experience. If the brain worked like *that*, it would make sense that experience feels like *this*.[26] In ordinary experience, an overall pattern of wakeful activity is modulated by many influences, with many network processes are tied together, in a way gives rise to the shifting whole of an experiential profile.

---

switching process could not be faster than 6-7 times per second (roughly a "seven-Mississippi" rate of change).

[25] Dehane and Naccache, "Does Claude possess a conscious global workspace?"

> A second signature [of the neuronal workspace] is limited capacity: the workspace acts as a bottleneck that can only attend to one representation at a time. This property explains why attending to a given process prevents you from becoming aware of another (*inattentional blindness*, as in failing to see a person dressed up as a gorilla) or delaying its perception by hundreds of milliseconds (*psychological refractory period*).

Also:

> Conscious processing, according to this [workspace] view, is a *function*, the temporary selection of one piece of information and its global broadcasting to all receiving processors, so that any processor can read it and act on it.

[26] I'm not seeing this as an argument *for* the view of NS activity (not: we can read the character of NS activity back off this feature of experience. Maybe something like that could be entertained in principle, but I don't know which other kinds of NS activity would suffice..). The claim is rather: this is a way in which special features of NS activity – linked to the hardware but with functional consequences – could be important in the explanation of consciousness.

Ordinary experience arises from having a brain in which the sidewalk sights and sounds are being processed by networks, and their combination into a single "field" of experience is not achieved by their networks sending signals to an integration-station of some kind, but by the whole pattern of activity being unified by the electro-biology of the brain.

I'll return now to an objection that began to come onto the scene earlier. Let's assume, for discussion, that I am right about LSDs and the unity of experience in our case. Suppose this is how it's done in us. That would be a big thing, and I'm sure that a lot of people would not want to accept it, but let's assume it for now. Then: surely the task that oscillations perform in us could be done in other ways? Surely some other way of achieving synchrony could be used, beside oscillations, or some other way of achieving unification, beside synchrony, might be used? I assume here that both claims are true. It doesn't follow that these features of brains are mere details of the substrate, subcomputational minutiae, irrelevant to mentality. These are *functional* features of the brain; functional properties can be finer or coarser-grained. If you take those features of brains out of the picture, then you might well have a system that can navigate a sidewalk, but you will have a system that does it differently from us. It will have coarse-grained functional similarities to us, and a host of fine-grained functional differences.

I also want to emphasize the in-principle points mentioned at the start: if I am wrong about LSDs, this discussion can give us a sense of *how* biological and hardware-related features might matter. Those features have fine-grained functional consequences, and may be "fine-grained computational" (or fine-grained quasi-computational) in their role. Certainly the details of our experience, what it feels like to be a *human* on a busy sidewalk, could be affected by these features.

I imagine someone then saying: OK perhaps so, but that doesn't mean these biological features are essential to the presence of experience in *some* form. Maybe a system without this biology will feel different, but it still might feel something? I'll take up this theme in the final section, but there's another one to cover first.

5. Portability

Let's suppose that LSD processes are important to cognition and experience in us. How hard would it be to bring them into a machine?[27] If someone wanted to realize something close to distinctly human experience in a machine, and thought that LSDs of this kind matter, what sort of obstacle would this be?

First, I'll return to a deferred point. I more or less accepted that the network interactions that are usually seen as central in brains could be realized in a computer. Neural network talk is ubiquitous around AI. What status does this have? A computer is not physically organized as a network. It's organized in a way that has a vast number of memory locations (of various kinds), and the values of variables are read and modified in these memory locations by a much smaller number of processors. A person might then say: the status of "network" activity in a computer is wholly representational, and dependent on our interpretation. We interpret some collections of read-write activities as "this node in the network influences that node, with this weight." The node is not in a physical place, like a neuron is. The state of "this node" will be represented across many locations, and these can change.

If the status of a "network" in a computer is representational in this way, it seems that the same standard should be applied to LSDs. We might represent in the computer's memory a 3D map of a brain, with the electrical properties of each location in the brain tracked. An "LSD update rule" would modify the values of variables in memory in a way that kept track of how the LSD properties of this modeled brain were changing. The neurons (also represented) would have their responses to influences of various kinds affected by the (represented) LSD changes.

I disagree with this line of thought. First, the "node role" is physically fulfilled to some extent in computer neural networks. A node is just a set of memory locations, but they are *treated by the machine* in its operation in a way that gives them node-like properties. It is a thing that affects other things of the same kind, in the targeted ways seen in a network. This (scattered, transient) part of the machine is made to act in a node-like way, and that is not an interpreter-dependent fact (or not very). In the case of LSDs, we might certainly *learn* a lot about them with models of the kind sketched a moment ago, but this would not be a realization of those properties in the computer system. There's no reason to believe that the consequences of a brain's LSD processes

[27] I skipped most of this section in the talk, thinking it would be the least controversial section.

for experience would also arise in a model that merely represents what they do. Instead, to be a contender for realization, there would have to be an artifact in which parallel computational operations were tied together by some additional spanning process, linking them with respect to their physical character (not: taking their various results and putting them together in a new computational operation). I assume that this is possible, but (I take it) far from anything around now.

6. Other Issues

I'll wrap up. First, I am not a neuroscientist, but there seems to be empirical reason to think LSDs matter – the "hum of the machine" view is fading. LSDs might be replaceable with respect to their coarse-grained functional role, but this seems to be *how our brains do things*, and that might have a lot of consequences for how it feels to be us. The features and texture of human EPs might have a lot to do with the biology of our brains. If you want a computational artifact with something like our kind of experience, you need a hardware that gives rise to similar fine-grained functional features in that system– features not merely represented, but realized. I assume, again, that this is possible, but different from anything that exists now. (The situation might change quickly.)

Is this kind of biology, or a fine-grained functional analogue of it, necessary to having felt experience of any kind at all? Suppose there is a system with a wide range of subjectivity-related properties – perception and action, a body, non-arbitrary selfhood, a rich handling of reward... but a completely different hardware with no LSD-like mechanisms. I don't concede that such a being could be behaviorally indistinguishable from an ordinary human or other animal, but I assume it could be behaviorally competent. Does this system have a different kind of consciousness, or something other than consciousness but related to it?[28] It would have to have a different way of integrating the disparate computational activities (in networks or something else) that enable it to see and hear and so on.

[28] A person might say: "Well, the question is whether there's *something it's like* to be them. That is clear enough." I don't think it's clear enough. The Nagel "something it's like" is a good gesture, but it's only a gesture and has no diagnostic role. A system with a non-animal hardware can have a lot of subjectivity-related properties, and then there's a way (something) that things *seem* like to it. But perhaps there can be rich but unfelt subjectivity.

The view I've outlined might seem an odd one within which to say: if you lack the special feature I'm discussing, then you are unconscious, a zombie. This is because the missing factor seems somewhat "medium-sized." But suppose, as I suspect, that something like an experiential profile is constitutive of "felt subjectivity" or felt perspective. More specifically, suppose you think that perspective-related phenomena are essential to experience (cf. Nagel), and when we think we're imaging experience without these features, we're just imagining a very unusual point of view. And suppose again (I don't know how I could show this), that something like an experiential profile is essential to having a "felt perspective." Then something like an EP is not really optional. If something EP-like is essential to experience, then a system needs something that generates it. This might be something other than LSDs, but that "something" will give rise to distinctive fine-grained functional features, and differences from a system like us.

I emphasize one more time my more general point, as opposed to the suggestions about LSDs. Hardware differences may make for fine-grained functional differences that matter to the explanation of consciousness, and that are not portable into an ordinary computer system. I suggest the LSD situation is actually like that, but if not, there may be other biological (sense 1) features of this kind.

What about other animals? The neural features I've emphasized are all over the animal kingdom, in a way that still surprises me. I am not sure how tight to make the connection to this next point, but over the last decade and even the last few years, evidence on the behavioral side for felt states in some invertebrates has increased a good deal. Insects have gone from being a frontier case to no longer a frontier case, I would say. Standard methods used to investigate pain in animals like birds and fish have always been applied occasionally to invertebrates, but especially in the insect case, the picture didn't become clear for a while, and now it has begun to. The idea that insects have pain-like valenced states now looks at least fairly likely. The case for cephalopods, which was always stronger, is now more so. We find wound-tending in bees, crickets, cuttlefish, octopuses, and shrimp. In octopuses and cuttlefish, wound-tending behavior is modulated by analgesic drugs.[29] The frontier cases are now neurally

[29] Some of the recent work: Gibbons et al., "Noxious stimulation induces self-protective behavior in bumblebees." Kuo et al., "Impact of Lidocaine on Pain-Related Grooming in Cuttlefish." Manzi et al., "Flexible Self-Protection as Evidence of Pain-Like States in House Crickets."

simpler animals – gastropods (slugs and snails) for example, with much more limited behavior. At this frontier, or the next one, I think we again reach a situation where the language and distinctions we use to address these questions run out of resources, and need to be partially replaced with a finer grained framework, a framework with richer contrasts than the familiar one between "conscious" organisms and the negative "non-conscious" ones.

I'll add a thought about Origins (apt for the Gálapagos Islands) to conclude. Nervous systems might nearly always have had their combination of networked and LSD features, but when we look at non-neural living systems, and look generally at evolutionary pathways, we can see two kinds of road. There is the evolution of information processing in networks, and the evolution of modulated-LSD configurations. Put these together (as nervous systems did, and gene regulation didn't), turn up the volume both times, and perhaps you get the thing we find in ourselves and label as experience or consciousness. The claim is not: "everything with some of this combination is conscious; *Hydra* is conscious." To say: "this is the path" is not to say "the end of the path is there the whole way along."

Appendix on AI

An emphasis in this talk is the role of hardware, and the way that hardware differences give rise to fine-grained functional differences that might matter to consciousness. Eventually I reached this question: even if some biological$_1$ features are important to *our* kind of consciousness, and hard to bring into a computer, how could we work out whether features of that kind are necessary for *any* kind of consciousness? How do we handle the boundary between "another form of consciousness" and "something other than consciousness, but related to it"? A hard question. But with that in mind, I want to make some extra comments about the possibility of consciousness in current AI systems, especially the LLMs that elicit a strong impression of the presence of a conscious interlocutor in many people who interact with them. These systems don't *only* have a hardware problem of the kind discussed in the talk. Their overall organization – the way they do what they do – seems to me to be extremely unlikely to give rise to consciousness.

People have offered a lot of possible disqualifiers in this area, involving embodiment, sensing, and their overall organization. I want to focus on the relationship between how they do things, and the evidence suggesting they might be conscious. When I say "how they do things," I don't mean what's happening at the hardware level (though that is interesting – see Shiller's "How many digital minds can dance on the streaming multiprocessors of a GPU cluster?"), but the overall computational structure, described in a standard way.

An LLM (ChatGPT, Claude) is a feedforward network that runs through a sequence of operations to produce a new linguistic fragment that continues the sequence of linguistic fragments that the system was prompted with. Each new word or similar fragment, each "token," is produced by running the linguistic exchange that has taken place up to that point into the system as input, which induces the system to deliver the next token. The LLM then runs the sequence of tokens that includes its own just-produced one back through the network, and comes up with another new token, continuing this process until it finds itself producing an "end of sequence" token. There is no ongoing activity beside this – no rumination, no looping of activity between the production of linguistic outputs. Some aspects of their design are changing now, but plenty of people have been convinced that LLMs are conscious when they're running as I am describing them here – with just repeated flashes of forward-running activity that come up, each time, with a new token to add to the sequence. It would be OK to regard the argument here as restricted to systems that have this organization.

Nothing about such a system makes it seem conscious other than its convincing verbal behavior. (As Anil Seth notes, almost no one thinks that AlphaFold, which predicts protein structure rather than engaging in chatty conversations, is conscious, though it's a similar type of AI system). There is only the system's sheer complexity, and the remarkable verbal output. And in this case, we have a deflationary explanation of how this behavior comes to seem so convincing. We are dealing with a giant mathematical device that has been trained to detect patterns in prior human language use, in a way that enables it to come up with plausible next moves, over and over again. That is how it does what it does.

Putting the same point a bit differently: We encounter a convincing interlocutor, as we often do. One interpretation of this kind of language use, whenever we encounter it, is that it is expressive of an inner life. Another possibility is that it is produced in a way that's indirectly reflective of the inner lives of others. A huge corpus of pre-

existing expressions has been brought together and mathematically processed in a way that enables new expressions of the same kind to be mechanically generated in a natural-seeming way. When we look inside an LLM, that is what we find. We find that *all* the activity is a cascade of operations that makes use of patterns in previous text to generate the next token. If it's not doing that, it's doing nothing.

This is not a matter of the wrong hardware. We're finding a mechanism for the production of linguistic behavior that actually realizes a deflationary option that we could already imagine, in at least a vague way, or can readily recognize once it's put in the table. What we find when we look "under the hood" is deflationary in a similar way as the hypothetical system in Ned Block's 1981 "Psychologism and Behaviorism." Block imagined a computer system that can convincingly carry on a text-based conversation with a human for some fixed period, and hence pass a Turing test, because it has stored a vast number of plausible conversational exchanges in a look-up structure that covers every possible move that might come from the human side of the conversation. An LLM is not set up in a way similar to Block's hypothetical system, but when we look under the hood, we again find a mechanism that's deflationary of the impression of an inner life that was suggested by its linguistic behavior. "Ah, *that's* how it comes up with plausible responses to everything I say."

In defences of the possibility of LLM consciousness, people sometimes say: "Some of their important features are *emergent*; you can't see them up close. If you look closely at neurons, they don't look very promising either, as a basis for experience." In response: if you zoom out just a little from the neurons (as we are doing with the AI), I don't think that's true. Look at what those neurons are doing. They are running a body, dealing with needs, responding to many kinds of sensory input, guiding action, and have all kinds of activity going on when you're not saying anything. No amount of emergence-oriented squinting will locate those features in an LLM. (Again, all this is true before we get to the special hardware properties discussed above.)

One kind of "emergent" feature that people consider in the LLM case is a tacit model of the world. At a low level, they are pattern-recognition devices that generate a plausible next token, but they might also, in virtue of this, have a model of the world in some sense. I don't object to that; a great many behaviorally capable devices might reasonably be said, in virtue of what they do, to have a model of the world, or of some part of it. That is different from claiming that a system has felt experience of some kind, an inner life.