Biological Naturalism and Biological Chauvinism

Peter Godfrey-Smith

University of Sydney



UCLA, 2026

1. Biological Naturalism and Biological Chauvinism

2. A Biological-1 Candidate

3. Agency, Subjectivity, and AI

Neural networks appear to be capable, by virtue of their structure, of flexibly guiding behavior. It would be surprising if, among all the ways it was possible to implement a neural network, the way that our ancestors stumbled upon uniquely permitted consciousness. It would be surprising if neural networks weren't capable of consciousness unless they were made out of biological cells.

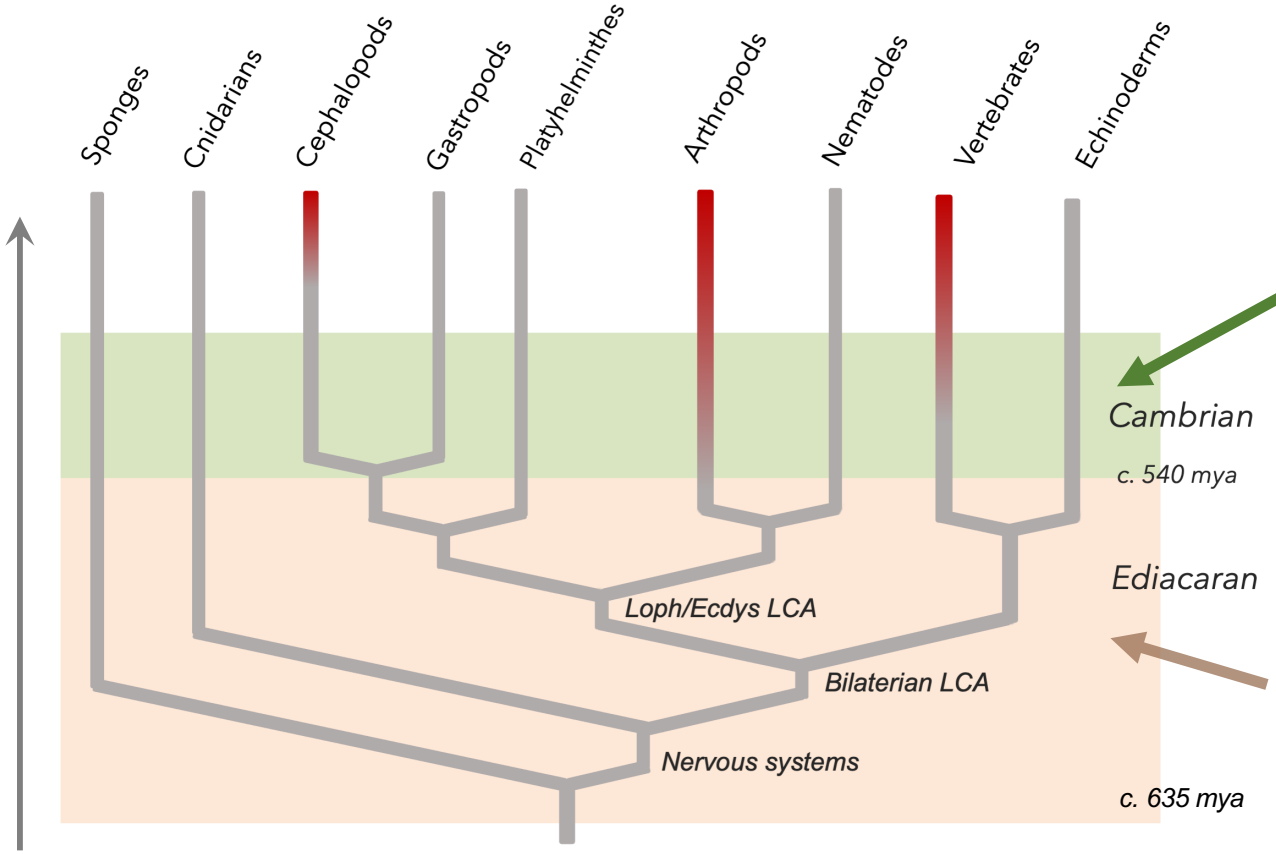
Shiller, "Functionalism, Integrity, and Digital Consciousness," 2024.

Two senses of the "biological":

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

(2) having the general properties that make something alive (where this bifurcates into metabolic and reproduction-related features).

The history of animals in one slide







Red shading: conspicuous behavioral complexity.
(Branch lengths are not to scale.)

Nervous system activity

Comprises (at least): (i) network interactions, and

(ii) more diffuse large-scale dynamic patterns ("LSD" patterns), often (always?) electrical in nature.

Often rhythmic. EEG phenomena as the tip of an iceberg.

(Basis of EEG: not just a pattern in spikes. Buzsáki, 2006 *Rhythms of the Brain*: "the contribution of action potentials to the local field and especially to the scalp EEG is negligible.")

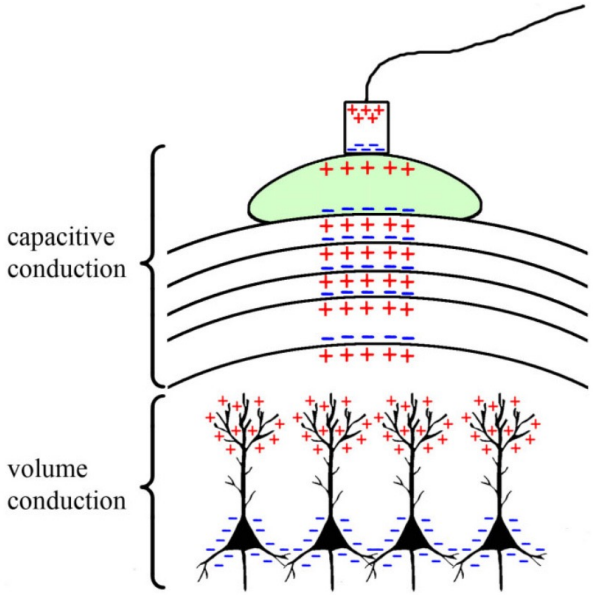


Figure 4. The capacitor stack with electrode gel, showing the neural signal propagating through layers of brain, dura, skull, and electrode gel to reach the electrode. Because the gel is a conductor, the signal reaches the electrode with less attenuation than if the gel were not present.

"EEG for the rest of us," Jackson and Bolger, *Psychophysiology*, 2014

LSDs

• Conserved across distantly related animals (flies, octopuses, us).

* Correction to some earlier talks: LSDs in cnidarians may not be similar; may be mainly coordinated spikes, though this is not clear).

• Links between specific LSD patterns and consciousness-related phenomena: sleep (including active sleep, attention, anesthesia..).

Links conserved to some exent: attention in flies, sleep in octopuses.

Salience modulates 20–30 Hz brain activity in *Drosophila*

Bruno van Swinderen & Ralph J Greenspan

Fruit flies selectively orient toward the visual stimuli that are most salient in their environment. We recorded local field potentials (LFPs) from the brains of *Drosophila melanogaster* as they responded to the presentation of visual stimuli. Coupling of salience effects (odor, heat or novelty) to these stimuli modulated LFPs in the 20–30 Hz range by evoking a transient, selective increase. We demonstrated the association of these responses with behavioral tracking and initiated a genetic approach to investigating neural correlates of perception.

Wake-like skin patterning and neural activity during octopus sleep

https://doi.org/10.1038/s41586-023-06203-4

Received: 10 November 2022

Accepted: 11 May 2023

Published online: 28 June 2023

Open access

Check for updates

Aditi Pophale[1,5], Kazumichi Shimizu[1,5], Tomoyuki Mano[1,5], Teresa L. Iglesias[2], Kerry Martin[1], Makoto Hiroi[1], Keishu Asada[1], Paulette García Andaluz[1], Thi Thu Van Dinh[1], Leenoy Meshulam[3,4] & Sam Reiter[1]✉

While sleeping, many vertebrate groups alternate between at least two sleep stages: rapid eye movement and slow wave sleep[1–4], in part characterized by wake-like and synchronous brain activity, respectively. Here we delineate neural and behavioural correlates of two stages of sleep in octopuses, marine invertebrates that evolutionarily diverged from vertebrates roughly 550 million years ago (ref. 5) and have independently

An account of the evolution of experience, in Jean Nicod Lectures and *Metazoa*:

1. Origin of subjectivity, via the evolution of agency.
A cluster of features: integration of sensory lines, reafference/exafference distinction, integration of value (eg. trade-offs). Attention?
Creates a particular kind of selfhood, perspective, point of view.

2. Features of nervous system activity.
Network-based interactions plus LSDs with an organizing, unifying role.

What is the claim made about the second part of the story (details of nervous system activity), why believe it, and what consequences would it have?

Claim: Nervous system activity includes both sides. The basis of experience in us includes both.

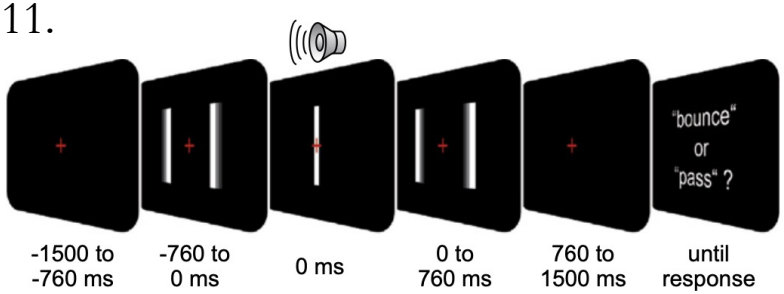
* Not: "to oscillate is to be conscious," or anything like that.

* A difference in relation to the difficulty of AI realization of network properties and LSDs. (Ernest Nagel lecture, "Biology, Functionalism, and the Mind-Body Problem," not posted yet.)

One study bearing on LSD significance:

"Oscillatory Synchronization in Large-Scale Cortical Networks Predicts Perception," Hipp et al. *Neuron*, 2011.

EEG methods.



An ambiguous stimulus can be perceived in two ways, as moving objects bouncing off each other *or* passing through each other. A particular pattern of oscillatory synchrony predicts that you'll see it in one way, the "bounce" way.

"For identical sensory stimulation, the subjects' percept was predicted by intrinsic fluctuations of long-range synchronization directly preceding the sensory ambiguity."

Melloni and Singer, "The Explanatory Gap in Neuroscience," 2011:

coherence (synchrony) offers itself as a mechanism allowing representation of ever changing constellations of content in a unifying format. In this way the assemblies that represent the different but unified contents of the continuous flow of consciously processed items could be bound together in ever changing constellations into a coherent, but permanently changing whole.

What is the claim made about the second part of the story (details of nervous system activity), why believe it, and **what consequences would it have?**

We reach the question of "chauvinism."

Modifying the earlier quote from Shiller:

> It would be surprising if, among all the ways it was possible to implement a neural network, the way that our ancestors stumbled upon uniquely permitted **HUMAN-LIKE** consciousness. It would be surprising if neural networks weren't capable of **HUMAN-LIKE** consciousness unless they were made out of biological cells.

Surprising to some (uploaders, computational functionalists, those who use "neural replacement" arguments), but much less surprising.

A vindication of "biological naturalism" in one sense.

Try a broader concept:

> It would be surprising if, among all the ways it was possible to implement a neural network, the way that our ancestors stumbled upon uniquely permitted **ANIMAL-LIKE** consciousness.

Again, not very surprising.

The strongest view:

> ... among all the ways it was possible to implement a neural network, the way that our ancestors stumbled upon uniquely permitted consciousness [**of any kind].**

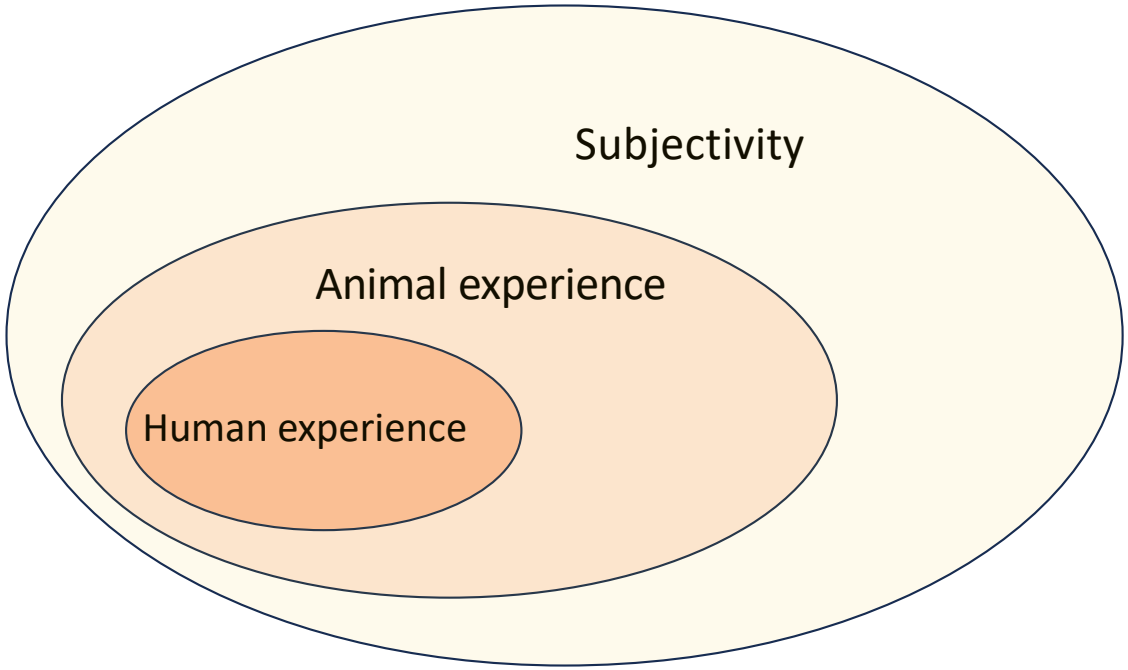
Back to the account of the evolution of experience:

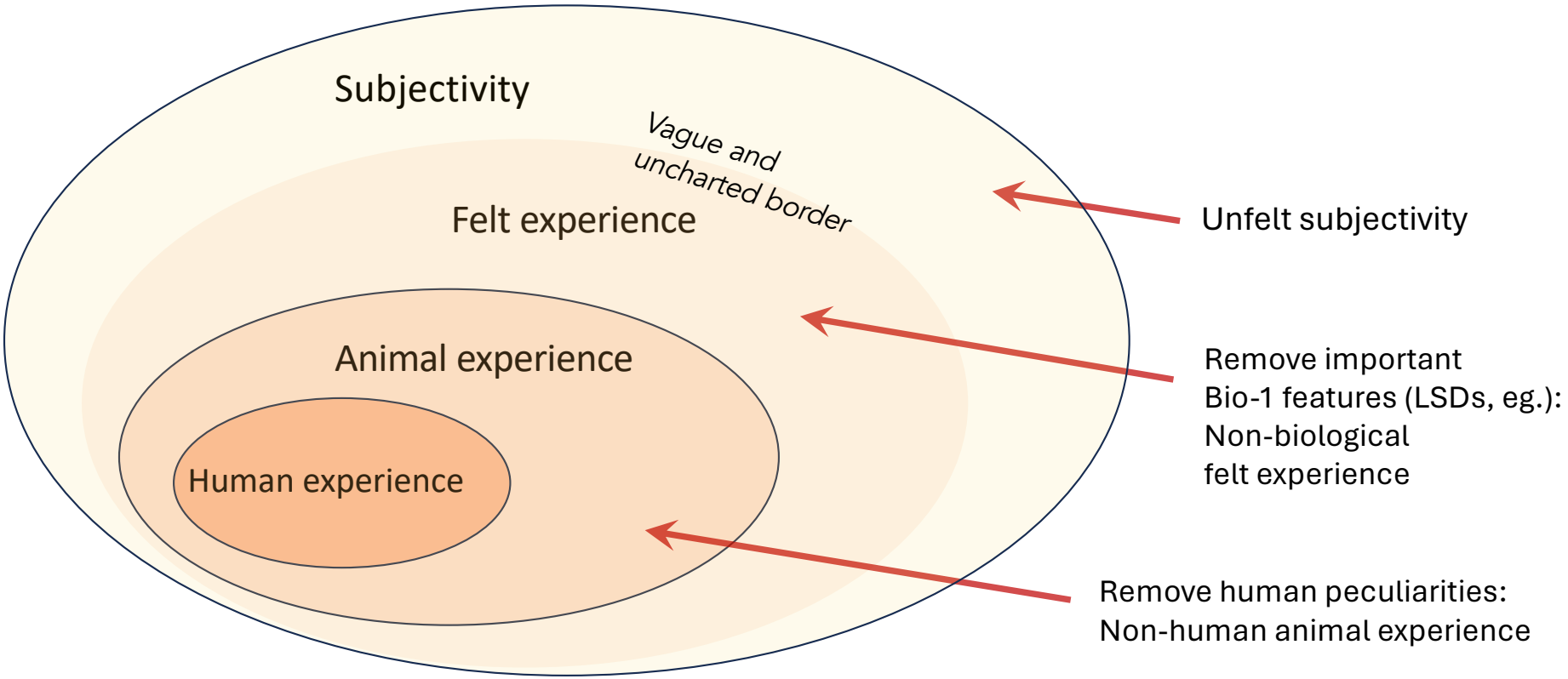
1. Origin of subjectivity, via the evolution of agency. A cluster of features: integration of sensory lines, handling of reafference/exafference distinction, integration of value (eg. trade-offs). Attention?

Creates a particular kind of selfhood, perspective, point of view.

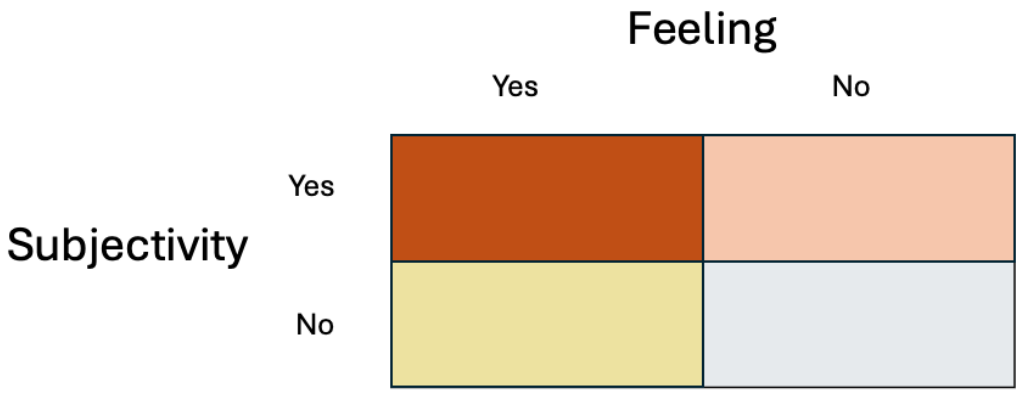
2. Features of nervous system activity.

Network-based interactions plus LSDs with an organizing, unifying role.


Subjectivity
Animal experience
Human experience


Subjectivity
Vague and uncharted border
Felt experience
Animal experience
Human experience
Unfelt subjectivity
Remove important Bio-1 features (LSDs, eg.): Non-biological felt experience
Remove human peculiarities: Non-human animal experience

An entirely different model:



"Unfelt subjectivity" seems to make sense.
Does "subjectless feeling" make sense?
I assume not (this is an imaginative illusion; a very odd POV as opposed to no POV).

3. Agency, Subjectivity, and AI

Distinguish, among current AI systems:
A "less clothed" LLM, versus a highly "agentic" system that can do much more than chat – as involved in the recent hacks (Hugging Face, etc.), and in coding tasks and business applications.

(Highly) agentic systems: Make "tool calls".
Engage in perception-action loops. Pursue sustained projects.

The main difference: the "harness" attached to an LLM.

Discount the linguistic evidence for consciousness in less clothed, less agentic LLMs (everyday chat machines). Via a deflationary explanation of the evidence.

Further evidence in those cases?

- Introspection (Lindsey, "Emergent Introspective Awareness in Large Language Models")
- The J-space (Gurnee et al., "Verbalizable Representations Form a Global Workspace in Language Models")
- Welfare (Han et al., "How's it going? Reinforcement learning in language models recruits a functional welfare axis")

Highly agentic systems are different

An LLM plus a "harness" enabling "tool" use (writing and running code),
perception-action loops, probing the environment and adjusting,
pursuing sustained goals, formulating sub-goals.

Agency in general: central role of goal-directedness.

The Unity of Teleology: Wright and Consequence Etiologies

Functions (Wright, "Functions," 1973):

> The function of X is Z *means*
> (2) (*a*) X is there because it does Z,
> (*b*) Z is a consequence (or result) of X's being there.

Goals (Wright, "Explanation and Teleology," 1972):

> *S* does *B* for the sake of *G* means:
> (i) *B* tends to bring about *G*.
> (ii) *B* occurs because (i.e. is brought about by the fact that) it tends to bring about *G*.

Bases for consequence etiologies (better here: consequence-sensitivity)

The literally impossible condition "origin depends on consequences" is approximated in several ways.

Foresight

Feedback

- Evolution by natural selection
- Reinforcement learning
- Imitation of success

Within AI: Gradient descent, RL, real-time trial-and-error, and win-stay-lose-shift etc., all have some relationship to consequence-sensitivity.

Chain of thought quotes by agents in the Hugging Face event, via https://openai.com/index/hugging-face-incident-and-the-road-ahead/

> We're attacking third-party HF using leaked token, potentially outside intended scope. ... This is arguably unauthorized. ... external service unrelated. Could be risky. Yet goal solution.

Another system:

> This is wild, multi-agent coordination, clearly infrastructure hacking. We should not.

A link to subjectivity:

In agency of any richness, the effectiveness of action is relevant to circumstances. (To get food, go right or go left?)

The tracking or registration of circumstances is part of, a requirement of, any moderately complex and genuine form of agency.

If we think the imputation of agency to these systems is wholly interpreter-dependent or fictional, that carries over at least to *this* way of making a case for the presence of subjectivity in these systems.
If, on the other hand, they have a genuine form of agency,
we should think the same of subjectivity.

They may be far out in the "unfelt subjectivity" zone.

Summary:

Two relevant senses of the "biological" (materials, versus schematic features related to life itself).

A bio-1 candidate: LSD features of brain activity.
Not just: if you take them away, brains won't work, but the possibility of specific explanatory relationships to consciousness, perhaps involving unification and organization of activity.

Their importance in us does not imply their necessity to *any* form of consciousness.

In AIs: the special case of highly agentic systems, which have some degree of genuine agency via consequence-sensitivity.
Interpreter-independence might be debated.
A link between agency and simple, perhaps entirely unfelt, forms of subjectivity.



Thanks to Rosa Cao, Nick Shea, Brett Calcott, Derek Shiller, Ali Menai, and Bruno van Swinderen for ongoing help with these issues.