10.20378/irb-109961

Defining Biological Naturalism

Johannes Kleiner[1,2,3]

Commentary on Seth, A. K. (2024). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1-42.

Preprint

[1]Institute for Psychology, University of Bamberg
Markusplatz 3, 96047 Bamberg, Germany
[2]Munich Center for Mathematical Philosophy, Ludwig Maximilian University of Munich
Geschwister-Scholl-Platz 1, 80539 München, Germany
[3]Association for Mathematical Consciousness Science
Markusplatz 3, 96047 Bamberg, Germany

Abstract:
Biological naturalism, as explicated by Seth, is indispensable for a balanced and metaphysically neutral science of AI consciousness. However, most of the properties that Seth explores might not be unique to living systems. Therefore, I argue, a biological naturalist research programme in consciousness science requires an explicit definition of Biological Naturalism. I discuss how such a definition might be obtained.

Main text:
Computational functionalism is well on its way to become the dominant—and perhaps also unquestioned—assumption across vast parts of consciousness science, especially where AI consciousness is concerned. Seth's piece is an important and indispensable counterweight to this trend. The intuitions and arguments he provides, based on scientific observations, are crucial to ensure a balanced and metaphysically neutral exploration of the AI consciousness question.

The next step in this exploration, then, should be the conception of biological naturalist theories of consciousness. Theories, that is, that provide an explicit and testable account of what consciousness is, which is true to "the idea that consciousness is a property of only, but not necessarily all, living systems" (p. 2).

One difficulty with this goal is that many of the properties of life that Seth discusses seem to be only contingently unique to living systems. It is hard to shake the feeling that many if not all of these properties could be instantiated by AI systems in the near future:

Predictive Processing and the Free Energy Principle, for example, are already employed to build next-generation AI systems (Friston et al., 2024). Some of the novel forms of analogue or neuromorphic computations, targeted at LLMs, utilize continuous-time dynamical processes and fine-grained timing relations (e.g. Schmidt et al., 2023). Mortal computation was proposed by Hinton (2022) with industrial applications in mind. Embodied, embedded, and quasi-agentic robots are already being constructed.

A biological naturalist theory of consciousness based on any of those choices is bound to bestow suitably designed AI systems with consciousness. And if substrate-dependent multiscale activities, generative entrenchment, or self-maintenance are of an evolutionary advantage, perhaps they are of a practical advantage in building AI systems as well. It is hard if not impossible to conceive of non-computational functional properties that cannot, theoretically, be implemented in synthetic systems that we would not otherwise conceive of as life.

This puts a biological naturalist research programme at a dilemma. It either has to operate with hypotheses that may all too easily attribute life to near-future artificial or synthetic systems, or it is relegated to a gap-filling strategy, where it is forced to invoke precisely those properties of life that are not, at a particular time, implemented in artificial systems.

This raises the question of whether it is possible to provide an explicit definition of Biological Naturalism that circumvents this dilemma: a definition that does not hinge on undefined notions such as life, but also avoids concrete hypotheses that are only contingently unique to living systems.

In the remainder of this comment, I would like to propose one way to obtain such a definition, based on a generalization of Putnam's original definition of Computational Functionalism (Putnam, 1967).

Crucially, Putnam's definition of Computational Functionalism is based on a "Description" (ibid., p. 54) of an organism in terms of Probabilistic Automata. It consists of four assumptions that explicate the idea that "being capable of feeling pain *is* possessing an appropriate kind of Functional Organization" (ibid.)—the Functional Organization specified by the Probabilistic Automaton Description.

Putnam makes use of Probabilistic Automata because they are the most general formal description of the type of digital general purpose computations available at his time,

Turing Machines with probabilistic transitions. However, from a mathematical perspective, nothing hinges on this choice of mathematical object. We can obtain a generalization of Computational Functionalism by allowing for *any* description of an organism in terms of a mathematical object. Referring to such a description as "Formal Description", this generalisation is:

Def. 1. *F*-Functionalism* is true iff:

1. Every organism capable of feeling pain possesses at least one Formal Description of a certain kind (i.e., being capable of feeling pain *is* possessing the kind of Organization described by the Formal Description).
2. No organism capable of feeling pain possesses a decomposition into parts which separately possess Formal Descriptions of the kind referred to in 1.
3. For every Formal Description of the kind referred to in 1, there exists a subset of the sensory inputs such that an organism with that Formal Description is in pain when and only when some of its sensory inputs are in that subset.

Because Probabilistic Automata are one type of Formal Description, *F*-Functionalism* provides a strict generalisation of Computational Functionalism. It also encompasses other forms of functionalism, to the extent that the structure of scientific theories can be represented in formal terms. And it does justice to novel developments in the computing industry, ranging from quantum computation to neuromorphic computation, that break the confines of Turing-type computation. The intuition is that being capable of feeling pain *is* possessing the kind of Organization described by the Formal Description. And equally so for other experiences.

Based on *F*-Functionalism*, Biological Naturalism can be defined in two different ways. A first strategy would be to invoke Formal Descriptions that involve mathematical objects that cannot be built—and hence not artificially or synthetically created. This is, essentially, endorsing the first horn of the dilemma mentioned above. In light of progress in modern physics, and in light of the surprisingly vast range of descriptors of computational systems (cf. e.g. (Aczel, 1988)), this strategy might not provide an easy option.

A second, easier, way is to define Biological Naturalism simply as the absence of a Formal Description as required by *F*-Functionalism* as follows:

Def. 2 *Biological Naturalism* is true iff there is no Formal Description that satisfies the three requirements of Definition 1.

Definition 2 singles out consciously experiencing organisms—conscious life—as systems that do not adhere to one common Formal Description. To paraphrase Seth, instead of a rigid organization, where consciousness is concerned, “there is an excitable fluidity to the matter of life” (p. 20). To the extent that the production of artificial or synthetic systems requires a formal description of the system that is produced, Definition 2 precludes artificial and synthetic systems from being conscious in virtue of what they are. Whatever makes something conscious, it is not a matter of having the right kind of Formal Description.

Acknowledgements:
I would like to thank Tim Ludwig, Lenore Blum, and Hanna Tolle for many insightful discussions on the topic of Computational Functionalism and AI consciousness,

Competing Interests Statement:
Competing interests: none.

Financial Support/Funding Statement:
This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

References:

Aczel, P. (1988). Non-Well-Founded Sets. *CSLI Lecture Notes*: Number 14. Stanford: CSLI Publications.

Friston, K. J., Ramstead, M. J., Kiefer, A. B., Tschantz, A., Buckley, C. L., Albarracin, M., ... & René, G. (2024). Designing ecosystems of intelligence from first principles. *Collective Intelligence*, 3(1), 26339137231222481.

Hinton, G. (2022). The forward-forward algorithm: Some preliminary investigations. *arXiv preprint* arXiv:2212.13345, 2(3), 5.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion.* Pittsburgh: University of Pittsburgh Press. (Reprinted in (Putnam, 1975). Page numbers reference the reprinted version.)

Putnam, H. (1975). The nature of mental states. In *Mind, language, and reality: Philosophical papers* (Vol. ii). Cambridge: Cambridge University Press.

Schmidt, H., Montes, J., Grübl, A., Güttler, M., Husmann, D., Ilmberger, J., ... & Schmitt, S. (2023). From clean room to machine room: Commissioning of the first-generation BrainScaleS wafer-scale neuromorphic system. N*euromorphic Computing and Engineering*, 3(3), 034013.