Commentary on: Anil Seth, “Conscious artificial intelligence and biological naturalism”

Word counts:

Abstract: 34
Main text: 870
References: 353
Entire document: 1,373

The question is not “are they conscious?” but “what could they be conscious of?”

Chris Fields[1] and James F. Glazebrook[2]

[1]Allen Discovery Center, Tufts University, Medford, MA 02155, USA

fieldsres@gmail.com; +33 6 44 20 68 69; https://chrisfieldsresearch.com
ORCID: 0000-0002-4812-0744

[2]Department of Mathematics and Computer Science, Eastern Illinois University, Charleston, IL 61920–3099, USA, and Adjunct Faculty (Mathematics), University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

jfglazebrook@eiu.edu; +1 217-328-4842; https://publish.illinois.edu/james-glazebrook/
ORCID: 0000-0001-8335-221X

Abstract: We argue that by focusing on an unanswerable ontological question, both computational functionalism and biological naturalism miss the empirical question of scientific and practical interest: what are systems of some kind *capable of* experiencing?

Anil Seth’s critique of computational functionalism as a theory of consciousness, and defense of biological naturalism as an alternative, are strikingly similar to those presented in *BBS* 45 year ago by John Searle (1980). “Minds, brains, and programs” was written in the depths of an AI winter, six years before the broad adoption of back-propagation (Rumelhart, Hinton and Williams, 1986). We are now in the midst of an AI boom, with achievements in AI considered worthy of Nobel Prizes. This suggests that something important has not changed, despite the enormous technical progress in AI, neuroscience, and the rest of biology in the past 50 years. We point to one thing that appears not to have changed: an assumption that whether something is computing is an *ontological* question.

Seth, like Searle, focuses on what he calls “Turing computation,” and again like Searle, links this conceptually with Turing computability. “Turing computation,” for both authors, is computation by von Neumann machines. Seth devotes considerable discussion to architectures that are not von Neumann, but neglects one that is arguably least von Neumann, that of reservoir computing (Tanaka et al., 2019). This focus on architectural differences belies an ontological assumption: that some things are “computers” with behavior that counts as “computation” while other things are not.

Whether a process is computation, however, is not a matter of fact, but a matter of measurement and interpretation (Horsman et al., 2014), and a given process can be interpreted as computing many different functions (Bongard and Levin, 2023). Measurements provide the data for interpretation, and all measurements have finite resolution; hence computational interpretations of even "analog" processes are always discrete (Fields, 1989; Gisin, 2021). The theory of Turing computability – for which the Turing Machine is just a convenient model – is the theory of functions "which could naturally be regarded as computable" (Turing, 1937, p. 230). Constructing a computational interpretation requires computing a function (Horsman et al., 2014), so what Turing computability constrains is our ability to interpret processes, including quantum processes (Deutsch, 1985), as computations.

When "computation" is seen as a fundamentally interpretative notion, the idea that dynamical systems, which are just physical processes defined on state spaces, are not "computing" ceases to make sense. Viewing the Free Energy Principle (FEP), which describes dynamical systems, as making any kind of distinction between living and nonliving, or conscious and nonconscious, systems (Sect. 4.2) is, moreover, missing the point of the FEP (Fields, Glazebrook and Levin, 2021; Fields and Glazebrook, 2025). The FEP states the condition under which a system – any system – that has a Markov blanket will persist in having one. Whether a system meets this condition is a matter of measurement and interpretation; indeed whether a system is entangled with its environment, and hence whether it has a Markov blanket, is a matter of measurement and interpretation (Zanardi, 2001). The interpretation step is not optional -- no system can demonstrate by measurement alone that any system, including itself, is unentangled with its environment (Fields, Glazebrook and Levin, 2024).

All systems, therefore, can be regarded as a computers, and claims that a system exceeds the computational power of a Turing Machine – e.g. that it solves the Frame (McCarthy and Hayes, 1969) or Halting (Turing, 1937) problems – can be neither refuted or confirmed by experiments (Fields, Glazebrook, Marcianò and Zappala, 2025). The processes that we interpret as computations are, as Seth points out, what systems *do*. It is, however, only what systems *do*, what they report in some way, that we can observe. As Hawking (2002) points out, our observational abilities, and hence our measurements and interpretations, are all inevitably incomplete.

Searle missed this point about doing and observability, and Seth misses it here; so have many others in the past 45 years. What we can discover, by observation and experimentation, is what systems – including ourselves – are *capable of* being aware of, which is just what they are capable of differentially responding to (Fields, Glazebrook and Levin, 2021). Laptops are capable of being aware of key presses, and large language models (LLMs) are capable of being aware of text strings. Organisms – even bacteria – are capable of being aware of hundreds to many thousands of dimensions of variation in the states of their environments and bodies. Are laptops in fact aware of key presses – is there "something it is like" when a key is pressed? Suppose some rocks are not just capable of experiencing, but actually experience, nanosecond-scale pulses of light when the right kinds of cosmic rays pass through them. Does this matter? What does it change?

The key to the "real problem" of consciousness, in our view, is careful attention to what systems *could be* aware of. LLMs could be aware of text fragments, and maybe even of phrase-usage statistics, together with context. Rocks could be aware of heat, and bacteria quite likely are, in our view, aware of energetic and osmotic stress. None of these are anything like human consciousness.

We fully agree with Seth that humans thinking that AI systems have human-like awareness is both an actuality and a problem. Perhaps the antidote is a careful analysis of what AI systems *could be* aware of, setting aside the non-empirical, unanswerable question of whether they in fact are.

Competing Interests: The authors have no competing interests.

Funding Statement: The authors have received no funding towards this contribution.

References:

Bongard, J. and Levin, M. (2023) There's plenty of room right here: Biological systems as evolved, overloaded, multi-scale machines. Biomimetics 8, 110.

Deutsch, D. (1985) Quantum theory, the Church-Turing principle and the universal quantum computer. Proc. R. Soc. London A 400, 97-117.

Fields, C. (1989) Consequences of nonclassical measurement for the algorithmic description of continuous dynamical systems. J. Expt. Theor. Artif. Intell. 1, 171-178.

Fields, C. and Glazebrook, J. F. (2025) Distributed Information and Computation in Generic Quantum Systems. Springer, Cham, Switzerland.

Fields, C., Glazebrook, J. F. and Levin, M. (2021) Minimal physicalism as a scale-free substrate for cognition and consciousness. Neurosci. Cons. 7, niab013.

Fields, C., Glazebrook, J. F. and Levin, M. (2024) Principled limitations on self-representation for generic physical systems. Entropy 26, 194.

Fields, C., Glazebrook, J. F., Marcianò, A. and Zappala, E. (2025) Whether a quantum computation employs nonlocal resources is operationally undecidable. Preprint arxiv:2501.14298; in review.

Gisin, N. (2021) Indeterminism in physics, classical chaos and Bohmian mechanics: Are Real numbers really real? Erkenntnis 86, 1469-1481.

Hawking, S. (2002) Gödel and the End of Physics. Lecture at the Dirac Centennial Celebration. Centre for Mathematical Sciences, University of Cambridge, Cambridge, UK. Available online: https://www.damtp.cam.ac.uk/events/strings02/dirac/hawking.html (accessed on 13 January 2024).

Horsman, C., Stepney, S., Wagner, R.C. and Kendon, V. (2014) When does a physical system compute? Proc. R. Soc. A 470, 20140182.

McCarthy, J. and Hayes, P. J. (1969) Some philosophical problems from the standpoint of artificial intelligence. In: Michie, D. and Meltzer, B., eds., Machine Intelligence, Vol. 4. Edinburgh: Edinburgh University Press, pp. 463-502.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1986) Learning representations by back-propagating errors. Nature 323, 533-536.

Searle, J. R. (1980) Minds, brains, and programs. Behav. Brain Sci. 3, 417-424.

Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., Numata, H., Nakano, D. and Hirose, A. (2019) Recent advances in physical reservoir computing: A review. Neural Networks 115, 100-123.

Turing, A. (1937) On computable numbers, with an application to the Entscheidungsproblem. Proc. London Math. Soc. Ser. 2 42 (1937), 230-265.

Zanardi, P. (2001) Virtual quantum subsystems. Phys. Rev. Lett. 87, 077901.