Steam-engine naturalism

Invited commentary on Seth (2025) https://doi.org/10.1017/S0140525X25000032
(Accepted)

Manuel Baltieri [1,2] and Ryota Kanai [1]

[1] Araya Inc., Tokyo, Japan, and
[2] Department of Informatics, University of Sussex, Brighton, UK

manuel_baltieri@araya.org

Seth (2025)'s biological naturalism builds on ideas from predictive processing, the free energy principle and active inference to 1) argue that only (some) biological systems are in the consciousness business, and to 2) refute the possibility of artificial consciousness, i.e. consciousness in AI and other artificial systems. We find this perspective problematic for a few different reasons.

The starting point for biological naturalism is *predictive processing*. Predictive processing describes various computational and cognitive processes in the brain that can be modelled in terms of cycles of predictions and error correction at different spatio-temporal, and hierarchical scales (Clark, 2013; Hohwy, 2013). However, it appears that not much is inherently biological about it: while it can be applied to describe certain biological processes, the same appears to be true for several non-biological ones too. It has been shown, in fact, that the core ideas behind predictive processing come from standard signal processing theory (Spratling, 2016) and estimation theory, e.g. Kalman filtering (Rao, 1999) (see however (Baltieri & Isomura, 2021)), and are simply applied to describe neural and behavioural processes of different kinds. While it is hard to overstate the importance of mathematical models of brain processes, it is perhaps even harder to argue that predictive processing can explain the differences between a steam engine and a human brain (Baltieri et al., 2020), or a Braitenberg vehicle (Baltieri & Buckley, 2017). One could argue that Braitenberg vehicles (Shaikh & Rañó, 2020) and Watt governors (Van Gelder, 1995) have often been used to support the idea of simpler models of biological and cognitive systems, and their behaviours. However, using this to vindicate the proposal of biological naturalism would be a category mistake, similar to thinking of Markov blankets as describing physical boundaries in biological systems (Bruineberg et al., 2022b, 2022a).

A second motivating principle behind biological naturalism is then the *free energy principle* (Friston, 2010; Friston et al., 2023). According to the free energy principle, for a "thing" to exist it must appear as if it is minimising a variational free energy functional to perform approximate Bayesian inference on the states of the external environment that produce its

sensory inputs. This principle however doesn't come without problems (Aguilera et al., 2022; Biehl et al., 2021; Bruineberg et al., 2022b, 2022a; Di Paolo et al., 2022; Raja et al., 2021), and it is unclear once again what would make it useful to support biological naturalism. Over the years, the free energy principle has been proposed as a theory of brains first (Friston, 2010), of life and its origins next (Friston, 2013), and finally of "things" in more general physical terms, consistent with accounts from quantum, statistical and classical mechanics (Friston, 2019; Friston et al., 2023). It is thus surprising to see this proposed as useful to formalise biological naturalism: if the free energy principle is in fact correct, as a (re)description of some key aspects of the physical world, it can be used "to simulate and predict the sentient behaviour of a particle, person, artefact or agent (i.e., some 'thing')" (Friston et al., 2023). Thus if we were to take the free energy principle at face value, biological systems ought to be consistent with it much like a "particle" or an "artefact" ought to. Seth further argues that the free energy principle supports an explanation of autopoiesis, thought to be a central aspect of theories about the emergence of biological individuality and normativity (Maturana & Varela, 1980). It appears, however, that this principle does not in fact explain, model or even account for autopoiesis in biological systems, conflating levels of analysis and definitions of "autopoiesis" (implying self-creation, and a clear description of structure and organisation among other things) and "homeostasis" (Bruineberg et al., 2022a; Di Paolo et al., 2022; Nave, 2025; Raja et al., 2021; Suzuki et al., 2022).

Finally, Seth argues that *active inference* (Friston et al., 2017; Parr et al., 2022) seen from a control theoretic/cybernetic perspective well synergises with accounts of interoception, the sense of the body "from within". However, the cybernetic perspective proposed by Seth is hardly exclusive to biological systems, see the steam engine/Watt governor example already discussed above (Baltieri et al., 2020). Mathematically, this simply appears as a restatement of classical works on "good regulators" and its formalisation in the "internal model principle", which cover classes of systems that can be described as goal-pursuing by an external observer, and thus not only biological systems (Baltieri et al., 2025; Virgo et al., 2025). One of the leading examples of this internal model principle in particular is integral control (Wonham, 1976), a core part of modern control theory (Åström & Hägglund, 1995) with applications in biology (Yi et al., 2000) and neuroscience (Ritz et al., 2017), which can also be easily accounted for in active inference (Baltieri & Buckley, 2019), making the special role supposedly played by active inference for biological systems not clear.

Overall, nothing about predictive processing, the free energy principle and active inference seems exclusive to biological systems. In light of this, we thus believe that for a coherent theory of consciousness, we are left with one of the following choices:

1. Pursue Seth's biological naturalism, but with different tools. In other words, build an argument for biological naturalism on something other than predictive processing, the free energy principle and active inference. This is because none of these seem to be in the business of characterising *exclusively* biological systems, much less *some* biological systems, the ones that ought to display the necessary conditions for consciousness.
2. Reject Seth's biological naturalism. Specifically, abandon the idea that there's something inherently special about consciousness in biological organisms, fully

embracing *universality* (Kanai & Fujisawa, 2024) beyond biology, and frameworks that don't a priori exclude the possibility of consciousness in artificial systems, like the free energy principle and active inference.

3. Accept parts of Seth's biological naturalism, but aim for *biological universality*. This means settling for a notion of universality that applies only to biological systems. The aim would be to make biological naturalism "as universal as possible" by providing formal, mathematical criteria for what should be understood as "biological", e.g. what counts as a "biological neuron" if we assume that these units are somehow important for consciousness, or what counts as "being alive" if this ought to be the "minimal 'ground state' of conscious experience" (Seth, 2025).

Acknowledgments.
M.B. and R.K. were supported by JST, Moonshot R&D, Grant Number JPMJMS2012.

References

Aguilera, M., Millidge, B., Tschantz, A., & Buckley, C. L. (2022). How particular is the physics of the free energy principle? *Physics of Life Reviews*, *40*, 24–50. https://doi.org/10.1016/j.plrev.2021.11.001

Åström, K. J., & Hägglund, T. (1995). *PID controllers: Theory, design, and tuning* (2. ed). Instrument Society of America.

Baltieri, M., Biehl, M., Capucci, M., & Virgo, N. (2025). *A Bayesian Interpretation of the Internal Model Principle* (No. arXiv:2503.00511). arXiv. https://doi.org/10.48550/arXiv.2503.00511

Baltieri, M., & Buckley, C. L. (2017). An active inference implementation of phototaxis. *European Conference on Artificial Life 2017*, 36–43. https://direct.mit.edu/isal/proceedings-abstract/ecal2017/29/36/99577

Baltieri, M., & Buckley, C. L. (2019). PID control as a process of active inference with linear generative models. *Entropy*, *21*(3), 257.

Baltieri, M., Buckley, C. L., & Bruineberg, J. (2020). Predictions in the eye of the beholder: An active inference account of Watt governors. *The 2020 Conference on Artificial Life*, 121–129. https://doi.org/10.1162/isal_a_00288

Baltieri, M., & Isomura, T. (2021). *Kalman filters as the steady-state solution of gradient descent on variational free energy* (No. arXiv:2111.10530). arXiv. http://arxiv.org/abs/2111.10530

Biehl, M., Pollock, F. A., & Kanai, R. (2021). A Technical Critique of Some Parts of the Free Energy Principle. *Entropy*, *23*(3), 293. https://doi.org/10.3390/e23030293

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022a). The Emperor Is Naked: Replies to commentaries on the target article. *Behavioral and Brain Sciences*, *45*.

Bruineberg, J., Dołęga, K., Dewhurst, J., & Baltieri, M. (2022b). The Emperor's New Markov Blankets. *Behavioral and Brain Sciences*, *45*, e183.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Di Paolo, E., Thompson, E., & Beer, R. D. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences*.

Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138.

Friston, K. J. (2013). Life as we know it. *Journal of the Royal Society Interface*, *10*(86), 20130475.
Friston, K. J. (2019). *A free energy principle for a particular physics* (No. arXiv:1906.10184). arXiv. https://doi.org/10.48550/arXiv.1906.10184
Friston, K. J., Da Costa, L., Sajid, N., Heins, C., Ueltzhöffer, K., Pavliotis, G. A., & Parr, T. (2023). The free energy principle made simpler but not too simple. *Physics Reports*, *1024*, 1–29. https://doi.org/10.1016/j.physrep.2023.07.001
Friston, K. J., Fitzgerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active Inference: A Process Theory. *Neural Computation*, *29*, 1–49. https://doi.org/10.1162/NECO_a_00912
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Kanai, R., & Fujisawa, I. (2024). Toward a universal theory of consciousness. *Neuroscience of Consciousness*, *2024*(1), niae022. https://doi.org/10.1093/nc/niae022
Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition: The realization of the living*. Springer Science & Business Media.
Nave, K. (2025). *A Drive to Survive: The Free Energy Principle and the Meaning of Life*. MIT Press.
Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. MIT Press.
Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews*, *39*, 49–72. https://doi.org/10.1016/j.plrev.2021.09.001
Rao, R. P. N. (1999). An optimal estimation approach to visual perception and learning. *Vision Research*, *39*, 1963–1989.
Ritz, H., Nassar, M. R., Frank, M. J., & Shenhav, A. (2017). A control theoretic model of adaptive learning in dynamic environments. *Journal of Cognitive Neuroscience*. https://doi.org/10.1162/jocn_a_01289
Shaikh, D., & Rañó, I. (2020). Braitenberg Vehicles as Computational Tools for Research in Neuroscience. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.565963
Spratling, M. W. (2016). A review of predictive coding algorithms. *Brain and Cognition*. https://doi.org/10.1016/j.bandc.2015.11.003
Suzuki, K., Miyahara, K., & Miyazono, K. (2022). Who tailors the blanket? *Behavioral and Brain Sciences*, *45*, e213. https://doi.org/10.1017/S0140525X22000206
Van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92*(7), 345–381.
Virgo, N., Biehl, M., Baltieri, M., & Capucci, M. (2025). A "good regulator theorem" for embodied agents. *(Under Review at ALIFE Conference)*.
Wonham, W. M. (1976). Towards an Abstract Internal Model Principle. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-6*(11), 735–740. https://doi.org/10.1109/TSMC.1976.4309444
Yi, T.-M., Huang, Y., Simon, M. I., & Doyle, J. (2000). Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proceedings of the National Academy of Sciences*, *97*(9), 4649–4653. https://doi.org/10.1073/pnas.97.9.4649