frontiers | Frontiers in Science

TYPE Frontiers in Science Viewpoint
PUBLISHED 28 February 2023
DOI 10.3389/fsci.2023.1147911

OPEN ACCESS

EDITED AND REVIEWED BY
Richard S. Frackowiak,
Swiss Federal Institute of Technology
Lausanne, Switzerland

*CORRESPONDENCE
Karl Friston
k.friston@ucl.ac.uk

RECEIVED 19 January 2023
ACCEPTED 29 January 2023
PUBLISHED 28 February 2023

CITATION
Friston K. The sentient organoid?
Front Sci (2023) 1:1147911.
doi: 10.3389/fsci.2023.1147911

The sentient organoid?

Karl Friston*

The Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, London, United Kingdom

KEYWORDS

organoids, artificial intelligence, self-organization, sentience, neuronal, neuromorphic

A Viewpoint on the Frontiers in Science Lead Article

Organoid intelligence (OI): the new frontier in biocomputing and intelligence-in-a-dish

Key points

- Brain organoids could represent the ultimate form of neuromorphic computing.
- Organoids could also function as experimental models in neurobiology, especially as models of active sensing.
- Organoids also have important potential as *in vitro* patients, providing a personalised experimental model to test pharmacological and genomic interventions.

Introduction

The review—or perhaps White Paper—by Smirnova et al. (1) offers an incredibly useful orientation to the emerging world of organoids and the exciting opportunities ahead. This viewpoint picks up on three cardinal themes as viewed through the lens of the free energy principle and active inference, namely: the potential for organoids as sentient artefacts with (artificial) generalised intelligence, as experimental models in neurobiology, and as *in vitro* patients.

Organoids as intelligent artefacts

This theme can be framed in terms of machine learning and engineered intelligence, i.e., the use of organoids to study sentient behaviour and active computers. Smirnova et al. start their overview by comparing current approaches in artificial intelligence and machine learning research with natural intelligence, noting the six orders of magnitude difference between *in silico* and *in vivo* computers (i.e., von Neumann architectures and functional brain architectures). These differences are expressed in terms of computational and thermodynamic efficiency, speaking to a change in the direction of travel for machine learning—a direction of travel that may be afforded by the introduction of organoids.

From a theoretical perspective, this can be usefully understood in terms of the physics of self-organisation. If one commits to the free energy principle, then one can describe the self-organisation of cells, organoids, and brains as minimising a bound on the log evidence or marginal likelihood of sensory inputs (2). The self-organisation comes into play when the cell, organoid, or brain actively samples or selects its inputs (a.k.a., active inference). But why inference? Here, inference speaks to the fact that—to define this kind of universal objective function—one needs a model that defines the likelihood of any sensory exchanges with the environment. This model is variously known as an internal, world, or generative model and is entailed by the structure and message-passing on the interior of the structure in question. This formalisation generalises things like reinforcement learning by absorbing preferred states into the prior preferences of the generative model (3). In this way, one can then use variational calculus to derive a lower bound on model evidence—known as the evidence lower bound (ELBO) in machine learning (4)—to describe the dynamics, plasticity, and morphogenesis as a gradient flow on variational free energy (5). So, why is this a useful formulation?

In the present setting, this objective or Lyapunov function has efficiency baked into it. This follows from the fact that model evidence can be expressed as accuracy minus complexity, where complexity corresponds to the divergence between prior and posterior representations (6, 7): namely, the degree of Bayesian belief updating associated with dynamics, plasticity, or morphogenesis. The complexity term is especially interesting here: it effectively scores the degree of departure from prior beliefs and the inverse efficiency of computation (in the spirit of both information theory and universal computation). Furthermore, *via* the Jarzynski equality, minimising complexity also minimises the thermodynamic cost (8–10). In short, the computational and thermodynamic efficiency of neuromorphic computation—under the free energy principle—are just two sides of the same coin and are emergent properties of a physics of sentience (11).

Having said this, the theoretical considerations above only place a lower bound on thermodynamic cost in the sense that belief updating —or message passing—on von Neumann architectures introduces a further cost, sometimes referred to as the 'memory wall' or von Neumann bottleneck (12, 13), read here as the inefficiency induced by reading and writing connectivity tensors into working (computer) memory. Practically, this can induce profound thermodynamic inefficiency, even in biomimetic computers whose connectivity (i.e., the message-passing structure) is not embodied or physically instantiated. This suggests that the next generation of 'computers' may turn to photonic, quantum, and neuromorphic computing (12, 14). Clearly, brain organoids are the ultimate neuromorphic computer.

Organoids in neurobiology

Another perspective on the utility of brain organoids concerns their potential as an experimental model for understanding how the brain works. In other words, organoids offer an empirical system that can be used to answer foundational questions about functional brain architectures and neuronal message passing (15). There are many interesting issues here. Smirnova et al. make the point (albeit implicitly) that all the questions, tools, and advances in contemporary neuroscience can now be deployed to understand how organoids work. But why would we want to do this?

One clear answer is that the opportunity for experimental intervention is much greater in organoids than in brains studied *in vivo*. However, to achieve a good model of the brain, one must establish some construct validity for the organoid model before any firm conclusions can be drawn or generalised from organoid to brain. A review of neuroscience over the past decades speaks to some foundational principles that one might expect to emerge in organoid research. For example, there are two principles of functional anatomy in the brain—functional specialisation and functional integration (16). The former speaks to the specialisation or segregation of selective responses to particular sensory inputs (and motor outputs). This principle suggests that the scaffolding or induction of segregated responses (perhaps *via* the use of assembloids) will be a key issue in organoid research, both in establishing functional specialisation and determining the situations under which it emerges developmentally, and through activity-dependent plasticity. From the perspective of the free energy principle, this kind of segregation emerges under what physicists would call a mean field approximation, namely, a factorisation of probabilistic representations that best explain sensory or observable data (4, 17–19). In turn, this speaks to the nature of generative models that underlie structured message passing. These generative models can be read as a modern-day version of the good regulator theorem (20, 21): any system that successfully controls its environment must be a sufficiently good model of that environment. Perhaps the most celebrated example here is the functional segregation into dorsal ('where') and ventral ('what') visual pathways in the human brain (22, 23). Theoretically, this is entirely predictable from the fact that knowing 'what' something does not tell you 'where' it is and vice versa. The ensuing conditional independence then underwrites the factorisation of posterior representations—which reduces their complexity—in accord with the maximisation of model evidence, which is sometimes known as self-evidencing (24).

Functional integration is just the statement that functional specialisation reflects a sparse message passing or connectivity architecture. Empirically, these architectures are characterised in terms of functional and effective connectivity. In the neurosciences, this is usually done using methods such as dynamic causal modelling to assess the recurrent but directed connectivity among neuronal populations (7). The sparsity of this connectivity defines the factorisation above and other key architectural or structural attributes that one might hope to see emerge in organoids, namely, hierarchical structures (25, 26) that usually go hand-in-hand with the separation of temporal scales (27, 28). But how does one infer this kind of connectome?

Perhaps one insight is that empirical neuroscience uses carefully designed studies to evoke functionally specialised responses and thereby assess or infer the connectivity among neuronal populations. One might argue that exactly the same approach will be mandated in organoid research. Strategically, this has an important implication. It means that one should be

complementing the exquisite and impressive technological developments—in organoid research—with consideration of experimental design; in other words, designing the right kind of experiments to estimate the internal architecture and belief updating in response to sensory perturbations. Interestingly, this places greater emphasis on eliciting 'smart data' from organoids in contrast to 'big data'.

A second pointer from neuroscience (that may be relevant for organoid research) is the pragmatic turn or enactive focus over the past decades (24, 29, 30). In other words, an appreciation that message passing and belief updating is embodied, situated, and, crucially, under the control of the brain—or organoid. This control is of course foundational for the good regulator theorem above (20). The implication here is that closed-loop experiments will involve an organoid acting or selecting its exchange with the external world (31) whether this be another organoid, a computer, or robotic actuators. The key thing here is that the organoid can act on its world in a way that has consequences for sensory inputs. The mediation of these consequences by the world is what gets installed into the generative model and, therefore, the structure and connectivity of the organoid. If this commitment to active inference as a description of sentient behaviour is correct, it raises a fundamental question: can organoids self organise to recognise causal structure in the world *de novo* or do they need some initial conditions or priors, e.g., epigenetic specification through an evolutionary process? Early evidence suggests that, at a minimal level, *in vitro* cultures can recognise causal structure and do so in an enactive context (15, 31). It will be fascinating to see to what extent more elaborate world or generative models emerge through structure learning (32–34) in the absence of genetically endowed anatomical scaffolds.

Organoids as *in vitro* patients

Smirnova et al. offer an excellent survey of the translational potential of organoids. They highlight the ability to perform *in vitro* neuropharmacology (and possibly disconnection or lesion experiments) in ways that would not be possible in real patients. Furthermore, one might imagine that having particular cell lines—from various neurological or psychiatric cohorts—would permit the kinds of invasive studies in 'personalised' organoids that would not be feasible *in vivo*. This is a potentially exciting avenue of research because many neuropsychiatric conditions can be cast as a form of functional dysconnection. In particular, most psychiatric disorders can be read as a particular form of synaptopathy, involving aberrant neuromodulation or neuroplasticity: e.g., ranging from epilepsy through autism to schizophrenia and Parkinson's disease (35–37). Having a validated organoid model of plasticity—and faster fluctuations in synaptic efficacy—would be invaluable, especially if differences can be systematically traced back to functional genomics *via* the mediating molecular biology at the cellular and synaptic levels.

A final perspective—afforded by the free energy principle—speaks to the nature of invasive experiments enabled by organoids: the free energy principle rests upon something called a Markov blanket that individuates something (e.g., a cell, organoid or brain) from everything else (38, 39). The Markov blanket constitutes a set of states that separates internal from external states. The separation rests upon sparse coupling in the sense of the underlying dynamics or statistical independencies in terms of the density dynamics self-organising things must evince. The key point here is that the internal states, interior to the Markov blanket, are unobservable from the outside unless one breaches the Markov blanket, as in invasive measurements in neuroscience.

This is a fundamental problem when imputing the internal architectures and dynamics of any self-organising system, and it leads to the need to infer what is going on in the interior based on measurements of blanket states. In the context of an organoid, the blanket states would constitute all the sensory inputs that mediate external stimuli and all the outputs that influence the 'world' mediating those sensory inputs. In this respect, organoid research faces exactly the same problem (e.g., the ill-posed inverse problem of reconstructing interior source activity from sensor measurements of electrophysiology). It was interesting to see in Smirnova et al. that much energy is focused on high-density recordings that do not interfere with the internal states of an organoid. One might imagine that in a few years' time, organoid researchers will deploy tools to reconstruct internal dynamics and connectivity based on techniques developed in neuroimaging. However, with careful scaffolding and microfluidics it sounds as though organoid researchers may be able to breach Markov blankets in a way that has never been possible before.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

KF is supported by funding from the Wellcome Centre for Human Neuroimaging (Ref: 205103/Z/16/Z), a Canada-UK Artificial Intelligence Initiative (Ref: ES/T01279X/1), and the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3). This research was funded in part by the Wellcome Trust [205103/Z/16/Z].

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Smirnova L, Caffo BS, Gracias DH, Huang Q, Morales Pantoja IE, Tang B, et al. Organoid intelligence (OI): the new frontier in biocomputing and intelligence in-a-dish. *Front Sci* (2023) 1:1017235. doi: 10.3389/fsci.2023.1017235

2. Ramstead MJD, Sakthivadivel DAR, Heins C, Koudahl M, Millidge B, Da Costa L, et al. On Bayesian mechanics: A physics of and by beliefs. (2022) 13–25. doi: 10.48550/arXiv.2205.11543

3. Tschantz A, Baltieri M, Seth AK, Buckley CL. "Scaling Active Inference". (2020) International Joint Conference on Neural Networks (IJCNN): Glasgow, UK. 1–8. doi: 10.1109/IJCNN48605.2020.9207382

4. Winn J, Bishop CM. Variational message passing. *J Mach Learn Res* (2005) 6:661–94.

5. Friston K, Levin M, Sengupta B, Pezzulo G. Knowing one's place: A free-energy approach to pattern regulation. *J R Soc Interface* (2015) 12:20141383. doi: 10.1098/rsif.2014.1383

6. Spiegelhalter DJ, Best NG, Carlin BR, van der Linde A. Bayesian Measures of model complexity and fit. *J R Stat Soc Ser B-Statistical Method* (2002) 64:583–616. doi: 10.1111/1467-9868.00353

7. Penny WD. Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage* (2012) 59:319–30. doi: 10.1016/j.neuroimage.2011.07.039

8. Landauer R. Irreversibility and heat generation in the computing process. *Ibm J Res Dev* (1961) 5:183–91. doi: 10.1147/rd.53.0183

9. Jarzynski C. Nonequilibrium equality for free energy differences. *Phys Rev Lett* (1997) 78:2690–3. doi: 10.1103/PhysRevLett.78.2690

10. Evans DJ. A non-equilibrium free energy theorem for deterministic systems. *Mol Phys* (2003) 101:1551–4.

11. Friston K, Da Costa L, Sakthivadivel DAR, Heins C, Pavliotis GA, Ramstead M, et al. Path integrals, particular kinds, and strange things. (2022) 30–1. doi: 10.48550/arXiv.2210.12761

12. Indiveri G, Liu SC. Memory and information processing in neuromorphic systems. *Proc IEEE* (2015) 103:1379–97. doi: 10.1109/JPROC.2015.2444094

13. Zou XQ, Xu S, Chen XM, Yan L, Han YH. Breaking the von Neumann bottleneck: architecture-level processing-in-memory technology. *Sci China-Inform Sci* (2021) 64(160404):1–10. doi: 10.1007/s11432-020-3227-1

14. Mead C. Neuromorphic electronic systems. *Proc IEEE* (1990) 78:1629–36. doi: 10.1109/5.58356

15. Isomura T, Friston K. *In vitro* neural networks minimise variational free energy. *Sci Rep* (2018) 8:16926. doi: 10.1038/s41598-018-35221-w

16. Zeki S. The ferrier lecture 1995 - behind the seen: The functional specialization of the brain in space and time. *Philos Trans Soc B-Biol Sci* (2005) 360:1145–83. doi: 10.1098/rstb.2005.1666

17. Beal MJ. Variational algorithms for approximate Bayesian inference. PhD. thesis, university college London (2003).

18. Dauwels J. On variational message passing on factor graphs. In: *2007 IEEE international symposium on information theory* (2007) RIKEN Brain Science Institute: Saitama, Japan. 2546–50.

19. Parr T, Sajid N, Friston KJ. Modules or mean-fields? *Entropy (Basel)* (2020) 22:552. doi: 10.3390/e22050552

20. Conant RC, Ashby WR. Every good regulator of a system must be a model of that system. *Int J Syst Sci* (1970) 1:89–97. doi: 10.1007/978-1-4899-0718-9_37

21. Ashby WR. *An introduction to cybernetics*. London: Methuen (1979).

22. Ungerleider LG, Haxby JV. 'What' and 'where' in the human brain. *Curr Opin Neurobiol* (1994) 4:157–65. doi: 10.1016/0959-4388(94)90066-3

23. Goodale MA, Westwood DA, Milner AD. Two distinct modes of control for object-directed action. *Prog Brain Res* (2004) 144:131–44. doi: 10.1016/S0079-6123(03)14409-3

24. Hohwy J. The self-evidencing brain. *Nous* (2016) 50:259–85. doi: 10.1111/nous.12062

25. Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* (1991) 1:1–47. doi: 10.1093/cercor/1.1.1

26. Hilgetag CC, O'Neill MA, Young MP. Hierarchical organization of macaque and cat cortical sensory systems explored with a novel network processor. *Philos Trans R Soc London Ser B Biol Sci* (2000) 355:71–89. doi: 10.1098/rstb.2000.0550

27. Kiebel SJ, Daunizeau J, Friston KJ. A hierarchy of time-scales and the brain. *PloS Comput Biol* (2008) 4:e1000209. doi: 10.1371/journal.pcbi.1000209

28. George D, Hawkins J. Towards a mathematical theory of cortical micro-circuits. *PloS Comput Biol* (2009) 5:e1000532. doi: 10.1371/journal.pcbi.1000532

29. Thompson E, Varela FJ. Radical embodiment: neural dynamics and consciousness. *Trends Cognit Sci* (2001) 5:418–25. doi: 10.1016/s1364-6613(00)01750-2

30. De Jaegher H, Di Paolo E. Participatory sense-making. *Phenomenol Cogn Sci* (2007) 6:485–507. doi: 10.1007/s11097-007-9076-9

31. Kagan BJ, Kitchen AC, Tran NT, Habibollahi F, Khajehnejad M, Parker BJ, et al. In vitro neurons learn and exhibit sentience when embodied in a simulated game-world. *Neuron* (2022). doi: 10.1101/2021.12.02.471005

32. Tenenbaum JB, Kemp C, Griffiths TL, Goodman ND. How to grow a mind: statistics, structure, and abstraction. *Science* (2011) 331:1279–85. doi: 10.1126/science.1192788

33. Tervo DGR, Tenenbaum JB, Gershman SJ. Toward the neural implementation of structure learning. *Curr Opin Neurobiol* (2016) 37:99–105. doi: 10.1016/j.conb.2016.01.014

34. Gershman SJ. Predicting the past, remembering the future. *Curr Opin Behav Sci* (2017) 17:7–13. doi: 10.1016/j.cobeha.2017.05.025

35. Pellicano E, Burr D. When the world becomes 'too real': a Bayesian explanation of autistic perception. *Trends Cogn Sci* (2012) 16:504–10. doi: 10.1016/j.tics.2012.08.009

36. Powers AR, Mathys C, Corlett PR. Pavlovian conditioning-induced hallucinations result from overweighting of perceptual priors. *Science* (2017) 357:596–600. doi: 10.1126/science.aan3458

37. Obi-Nagata K, Temma Y, Hayashi-Takagi A. Synaptic functions and their disruption in schizophrenia: From clinical evidence to synaptic optogenetics in an animal model. *Proc Jpn Acad Ser B Phys Biol Sci* (2019) 95:179–97. doi: 10.2183/pjab.95.014

38. Kirchhoff M, Parr T, Palacios E, Friston K, Kiverstein J. The Markov blankets of life: autonomy, active inference and the free energy principle. *J R Soc Interface* (2018) 15:20170792. doi: 10.1098/rsif.2017.0792

39. Sakthivadivel DAR. Weak Markov blankets in high-dimensional, sparsely-coupled random dynamical systems. (2022) 15–6. doi: 10.48550/arXiv.2207.07620