

Foundations of Computation

SYMPOSIUM FOR THE AISB CONVENTION 2026

THE SOCIETY FOR THE STUDY OF ARTIFICIAL INTELLIGENCE
AND SIMULATION OF BEHAVIOUR



FERNANDO RODRIGUEZ-VERGARA, ADAM ROSTOWSKI & BEN WHITE
(EDITORS)



Foreword from the Symposium organisers

For the better part of a century, computation has supplied both the vocabulary and the explanatory currency of the sciences of mind. The settlement has been remarkably durable: to explain a cognitive capacity was to specify the computations that realise it, and disputes concerned which computations rather than the computational foundation itself. That settlement is now under pressure from an unexpected direction. The recent performance of large-scale artificial systems has not resolved the old foundational questions so much as made them newly urgent. Recent progress calls for a reassessment of some long-standing assumptions within both our practical and theoretical approach to computation.

This symposium focused on the ‘Foundations of Computation’ (FOCSY), inviting our participants to join in a reassessment of the settlement around computation, rather than attempting to raise any cause for alarm. The symposium addressed fundamental questions about the nature, applications, and limits of computation, with a focus on their bearing on cognition and intelligent behaviour in living and artificial systems alike. What is it for a physical system to compute, and who gets to decide? What does a computational explanation deliver that other explanatory strategies do not? What work are mechanisms and representations still doing in our models, and what would be lost, or gained, by doing without them? These questions are not new, but the conditions under which we ask them have changed enough to warrant asking again.

The symposium is deliberately positioned within AISB’s long-standing identity as a venue where artificial intelligence and cognitive science meet on equal terms. This meeting is only possible through the thoughtful, sustained, and high-quality contributions we received; the two communities and research traditions have accumulated separate literatures, separate benchmarks for what counts as progress, and separate intuitions about which foundational questions are live. FOCSY aims to provide a structured forum in which those differences can be made explicit and argued over, and to open space for approaches that fall between the established programmes — computational phenomenology among them, as a case where the vocabularies of first-person experience and of formal modelling have only begun to be brought into contact.

The contributions collected here reflect that spread. A first cluster presses on what computation is and where its boundaries lie. Michael Woodley approaches the question from mathematical and physical perspectives; Simon Bowes defends a computational functionalism grounded in the relation between concrete computations and virtual machines; Avel Güenin-Carlut contrasts the organisation of information in mechanical and biological systems, arguing that autonomy and contextuality mark a genuine difference between them; and Steve Battle takes up autopoiesis as a constructive problem rather than a metaphor. Kathryn Nave sharpens the challenge from the other direction, arguing from the principle of excessive variety that organic activity is not mechanical in character. A second cluster concerns

what computational explanation actually delivers. Alina Gutoreva asks what computation does and does not explain about everyday cognitive activity, and Oriol Roca-Martin raises a philosophical difficulty for emergent multi-scale causality as a modelling framework — a challenge that meets Fernando Rosas’s account of hierarchical emergence and multi-scale computation directly. A third takes up meaning and mind. Nathaniel Virgo proposes a framework linking computation to meaning through a duality between intrinsic and extrinsic goals; Dean Petters and Achim Jung argue that computational expressivity is the key to resolving Searle’s Chinese Room; and Ron Chrisley argues that even granting humans capacities no Turing machine possesses, digital computational consciousness remains a live possibility. Taken together they do not converge on a position, which is as it should be for a symposium whose premise is that the foundations remain open.

We are grateful to our keynote speakers: Kathryn Nave, Ron Chrisley, Nathaniel Virgo and Fernando Rosas for anchoring the day’s three sessions and setting the terms of the closing roundtable; to our reviewers, whose careful and timely work shaped the volume considerably; and to the AISB for hosting FOCSY as part of the 2026 Convention at the University of Sussex, and for its continued support of venues in which foundational questions are treated as research problems rather than as preliminaries to be dispatched.

Table of Contents

Mathematical and physical perspectives on the nature of computation <i>Michael Woodley</i>	1
An Overlooked Challenge for the “Emergent Multi-Scale Causality” Framework for Modelling Complex Systems <i>Oriol Roca-Martin</i>	3
Concrete computations and virtual machines: In defence of a computational functionalism <i>Simon Bowes</i>	6
Even if humans can answer questions Turing machines cannot, that doesn’t mean that digital computational consciousness is impossible <i>Ron Chrisley</i>	9
What Computation Explains, and What It Does Not about Daily Cognitive Activity? <i>Gutoreva, A. and Omar, Z. and Bektemisov, N. and Nurlankyzy, D.</i>	10
Computational Expressivity is the key to resolving Searle’s Chinese Room argument <i>Dean Petters and Achim Jung</i>	13
The relationship between goals and beliefs, and some implications for computation <i>Nathaniel Virgo</i>	16
Artificial Autopoiesis <i>Steve Battle</i>	20
Comparing the organization of information in mechanical and bio- logical systems <i>Avel Guénin–Carlut</i>	24
The Principle of Excessive Variety: The Brain as an Intrinsically Exploratory System <i>Kathryn Nave</i>	28

Mathematical and physical perspectives on the nature of computation

M. T. M. Woodley¹

1 Abstract

A broad overview is given of several different aspects of computation. The intention is that this overview complements other, perhaps more specialised presentations of related topics by describing computation from both mathematical and physical points-of-view. Particular emphasis will be given to highlighting some key connections, both to more recent developments in unconventional computing architectures, and to more historical developments in certain parts of the foundations of mathematics from the early 20th Century, especially with reference to the work of L. E. J. Brouwer and J. von Neumann.

2 Physical perspectives

At the time of writing, there is rapidly growing interest in unconventional approaches to computing – i.e. those that deviate significantly from the standard von Neumann architecture. An important reason for this lies in the separation between CPU and memory in this architecture [1], which can lead to bottlenecks in data throughput that are becoming increasingly severe and performance-limiting – particularly for machine learning applications. One type of unconventional approach is neuromorphic computing [2] – computing inspired by the structure and function of the human brain. In this case, high parallelism and processing in memory are both strongly emphasised, in order to more efficiently implement algorithms for which these characteristics are advantageous – particularly in machine learning.

In addition to the overall architecture of a computer, another important topic is the choice of physical substrate used for the computation itself. Several approaches have been suggested in addition to the conventional method of using electronics, including photonic and spintronic systems [2]. Photonic approaches have a particular advantage in relation to the parallelism of neuromorphic architectures, namely wavelength-division multiplexing [3]. This involves performing parallel processing using multiple wavelengths within a single optical channel or circuit, which significantly reduces the number of individual interconnects required in such a system.

Another physical aspect of computation is the computer as analogy for certain parts of the physical world. An important example of this is the wind tunnel as an analogue computer in the context of fluid dynamics [4]. Here, measurements are taken from a scale model of a structure or vehicle, subject to air flow. When certain scaling laws are approximately satisfied, this model acts as a simulation of the structure or vehicle in question. In this sense, computation is performed directly by the physics governing the interactions between the scale

model and the air flow; this serves as a simulation of the full-scale equivalents of those interactions.

The unifying theme of these physical perspectives, primarily on unconventional computing, is the tailoring of a physical architecture or substrate for the efficient realisation of certain approaches to computation.

3 Mathematical perspectives

From the mathematical point-of-view, it is instructive to emphasise an historically significant connection between computation and the philosophy of mathematics – specifically intuitionism. Intuitionism is an approach, introduced by L. E. J. Brouwer [5], that views mathematics as being the result of constructive mental activity, such that certain principles of reasoning, which otherwise feature in classical logic, are not taken to be valid – in particular, the law of the excluded middle (the assertion that, for any given logical proposition, either it, or its negation, must be true). Time plays a fundamental role in intuitionism, as a temporal sequence is considered to be the mechanism whereby an individual constructs distinct entities corresponding to the positive integers. This renders intuitionism significantly different to classical foundational treatments of mathematics. Also of interest in connection with intuitionism is the work of H. Weyl [6], whose support for this approach varied significantly during the early 20th Century.

A point of connection between intuitionism and computation can be seen via the concept of constructability. Addressing the realisability of intuitionistic logic, S. Kleene interpreted constructability in terms of computable functions [7] – i.e., those functions the value of which may be obtained by an effective method. The concept of time appears in the definition of an effective method as a finite-time, deterministic procedure. The connection between proofs and computer programs was subsequently formalised by the Curry-Howard correspondence [8]. Moreover, deep connections between intuitionistic logic, category theory, and functional programming have been established [8,9,10]. However, a detailed examination of these topics is outside the scope of this paper.

A fundamental model of computation is the theory of automata – the study of self-propelling abstract machines. An observation that may be made of an automaton is that its evolution takes the form of state transitions that implicitly make reference to the passage of time in the sense of an ordered sequence of steps. Moreover, it was observed by von Neumann – with reference to previous work by A. M. Turing, W. S. McCulloch, and W. Pitts – that intuitionistic forms of logic may be studied using the theory of automata [11]. In this way, a bridge may be appreciated between intuitionism, in which the passage of time is explicit and foundational, and descriptions of

¹ University of Bath, England
mw2970@bath.ac.uk

computation in which time does not explicitly feature, but which may be implicit in the concepts employed by those descriptions.

Finally, it is instructive to highlight more recent mathematical efforts in establishing the boundaries of computation. One fruitful example of this is work on the development and applications of the solvability complexity index (SCI) hierarchy, whereby classes of problems are characterised in terms of the smallest number of limits of sequences of approximations needed to compute solutions to those problems, given certain restrictions on the algorithm involved [12,13]. This development may be seen as a connecting piece that extends the theory of computation by classifying problems of particular interest in scientific computing, such as the computation of spectra, offering rigorous results on algorithmic convergence and error control, as well as proving fundamental barriers to algorithmic tractability. In this sense, it is complementary to other theoretical descriptions of computation and its fundamental limits, such as automata.

4 Conclusions

The primary goal of this paper has been to present several mathematical and physical perspectives of computation. This included the practical importance of tailoring the architecture and substrate of a computer for efficient implementation of certain types of computation. Attention was then given to mathematical intuitionism and its connections with computation. In particular, the concept of time was addressed, as this features explicitly in intuitionism. It was argued that the passage of time appears ‘in the background’ in certain concepts and models of computation, even if this is implicit. These included the definition of an effective method in the context of constructability, and the evolution of automata.

A view is offered here that computation may, in an important sense, be treated as the modelling and realisation of translating between deductive inference and cause and effect. In the light of various other types of unconventional computers that are being developed at the time of writing – including those that have not been addressed here, such as quantum computers – a potentially interesting generalisation of this principle for future work may be to rather consider computation in terms of the modelling and realisation of translating between inference and physics.

REFERENCES

[1] D. Kimovski, N. Saurabh, M. Jansen, A. Aral, A. Al-Dulaimy, A. B. Bondi, ‘Beyond Von Neumann in the Computing Continuum: Architectures, Applications, and Future Directions’, *IEEE Internet Computing* 28 (3), 6-16 (2024).

[2] T. Asai, ‘Novel Hardware for Unconventional Computing’, in ‘Unconventional Computing’, edited by A. Adamatzky, Springer New York, NY (2018).

[3] B. J. Shastri, A. N. Tait, T. F. de Lima, W. H. P. Pernice, H. Bhaskaran, C. D. Wright, and P. R. Prucnal, ‘Photonics for artificial intelligence and neuromorphic computing’, *Nature Photonics* 15, 102-114 (2021).

[4] D. A. Caughey, ‘Computational Aerodynamics’, *Encyclopedia of Physical Science and Technology*, Academic Press, Third Edition,

pp. 469-485 (2003).

[5] L. E. J. Brouwer, ‘Over de Grondslagen der Wiskunde (On the Foundations of Mathematics)’, Amsterdam-Leipzig: Maas & van Suchtelen (1907).

[6] H. Weyl, ‘Philosophy of Mathematics and Natural Science’, Princeton University Press (2009).

[7] S. C. Kleene, ‘Hierarchies of number-theoretic predicates’, *Bulletin of the American Mathematical Society* 61 (3.P1), pp. 193-213 (1955).

[8] W. A. Howard, ‘The formulae-as-types notion of construction’, in ‘To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus, and Formalism’, edited by J. P. Seldin and J. R. Hindley, New York: Academic Press (1980).

[9] J. Lambek, ‘From λ -calculus to Cartesian closed categories’, in ‘To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus, and Formalism’, edited by J. P. Seldin and J. R. Hindley, New York: Academic Press (1980).

[10] F. Cano-Jorge, ‘Extending computational trinitarianism’, *Logic Journal of the IGPL* 34 (4), jzag037 (2026).

[11] J. von Neumann, ‘Probabilistic logics and the synthesis of reliable organisms from unreliable components’, in ‘John von Neumann: Collected Works Volume V’, edited by A. H. Taub, Pergamon Press (1963).

[12] M. J. Colbrook, ‘The Foundations of Infinite-Dimensional Spectral Computations’ (PhD thesis), University of Cambridge (2020).

[13] C. Drysdale, M. Colbrook, and M. T. M. Woodley, ‘Computation and Verification of Spectra for Non-Hermitian Systems’, *Physical Review Letters* 135, 170202 (2025).

An Overlooked Challenge for the “Emergent Multi-Scale Causality” Framework for Modelling Complex Systems. A Brief Philosophical Note

Oriol Roca-Martín¹

Many of the current world challenges that scientists are trying to tackle are linked by a common thread: they involve the modelling, prediction, and control of complex systems, which are characterized by having properties that emerge from the non-trivially ordered interactions of large numbers of elements at different scales. Among the many significant challenging tasks of this kind are the research on climate change and its consequences, effective procedures and policies for mitigating unfair economic trends or global tax evasion schemes, the spread of misinformation and pandemics, or attractor dynamics in gene regulatory networks and the brain.

The classic assumptions of linearity, locally defined dependencies, and reduction have brought us only so far in effectively modelling the world; scientists studying complex systems on many occasions have to work, instead, with non-linearity (e.g., power laws, and phase transitions), strong path-dependencies, and emergent processes at different scales (for examples of the previous contrasts [27, 20, 17, 15]). We are far from a complete formal and conceptual framework that can systematically guide the modelling, understanding, and harnessing of complex processes. Taking into account these meta-theoretical needs, this article focuses on emergence.

Recently, various research strands on this topic have converged towards a set of closely connected ideas that make use of information-theoretic and computational complexity principles in order to account for and formally model emergence and causality across different scales; what I’ll call here the *emergent multi-scale causality* (eMSC) approaches. This framework is growing in popularity—in science [32] and in some prominent versions of ontic structural realism in philosophy [12]. However, the methodological and metaphysical considerations that underlie the modelling decisions of the framework remain unclear (as has been noted by numerous authors, e.g., [9, 8, 33, 3]). In the context of this concern, this article shows that the main strategic advantage of eMSC when it comes to defending scientific realism towards effective theories’ ontologies—that is, being able to systematically define indispensable ontologies (discussed below)—faces some problems. The reasons for these problems are articulated, and possible natural extensions of the framework are briefly discussed.

Despite significant (and interesting) differences in specificities and intellectual origins (which we will have to ignore here), four pillars indicate a common characteristic thread shared by the approaches that constitute the eMSC framework:

P1. Emergence is thought of in terms of the capacity to construct effective theories from observations of natural systems (Fig. 1a) (very similar ideas can be found in [20, 5, 13]). Given a task—solvable if a generalization (from H_1 to a target H_i among $H_1 \dots H_n$) robustly obtains with a fidelity down to some noise tolerance—, finding an ef-

fective theory means identifying the appropriate abstract or “coarse-grained” scale(s) containing variables and dynamics that robustly predict H_i , avoiding the costs (or the unsolvability) of tracking lower dynamics with more degrees of freedom. We can find tasks and effective theories that solve them “in the wild”, in natural sciences—a classic example of a scientific task can be tracking the position of Neptune, we can do so by following its centre of mass (following H dynamics given a generalization) in contrast to having to follow all the degrees of freedom of the particles that compose the planet (L dynamics in a particle-based generalization from another theory). And also in engineering—a digital calculator is cleverly crafted in such a way that it is a system in which L dynamics are robustly securing some specific H dynamics according to some generalizations that solve the tasks of, for example, adding and subtracting with a required fidelity.

P2. In the eMSC framework emergent entities are conceived as patterns compressed from regularities in datasets, and a “more inductive bang for the buck” realist metaphysics of science is adopted towards patterns: the metaphysical force comes from noticing that one cannot just wish into existence either facts about inductive success—a proposed effective ontology can either reproduce the required dynamics according to some noise tolerance, or it cannot—or facts about computational complexity—some successful effective ontologies are informationally more efficient than others (axis A and B in Fig. 1b). The framework adopts this stance from computational mechanics’ foundational way of reasoning about the topic [22], which is explicitly linked to Daniel Dennett’s Real Patterns ideas [7, 12]. (There are characteristic epistemic/evolutionary derivations of the previous metaphysical points, e.g., [16]).

P3. Just by following the previous criteria, however, a multiplicity of ontologies arises: indefinitely many successful compressions—each yielding its own ontology—can be obtained by systematically finding (even just marginally) different efficient ontologies for a given task. This is considered problematic for the realist aims of the framework, and raises the challenge of providing a scarcity criterion for decreasing ontological density [11, 14, 1].

P4. Faced with this situation, eMSC proposes an explicit scarcity criterion: indispensability. According to which the relevant effective taxonomy is the one that reproduces a generalization (with a fidelity down to some noise tolerance) at the lowest complexity costs (star in Fig. 1b). These four pillars have been independently identified and discussed by authors within philosophy of science [14, 26] and scientists modelling complex systems [22, 23, 31, 10].

But despite its clearly objective, mind-independent features (facts about inductive success and complexity-theoretic optimality) indispensability is not easy to completely cash out for the methodological and metaphysical debate at hand. The issue is that while indispensability might solve the Local Challenge of providing a scarcity cri-

¹ LOGOS, Barcelona Institute of Analytic Philosophy, Universitat de Barcelona; o.rocamartin@ub.edu

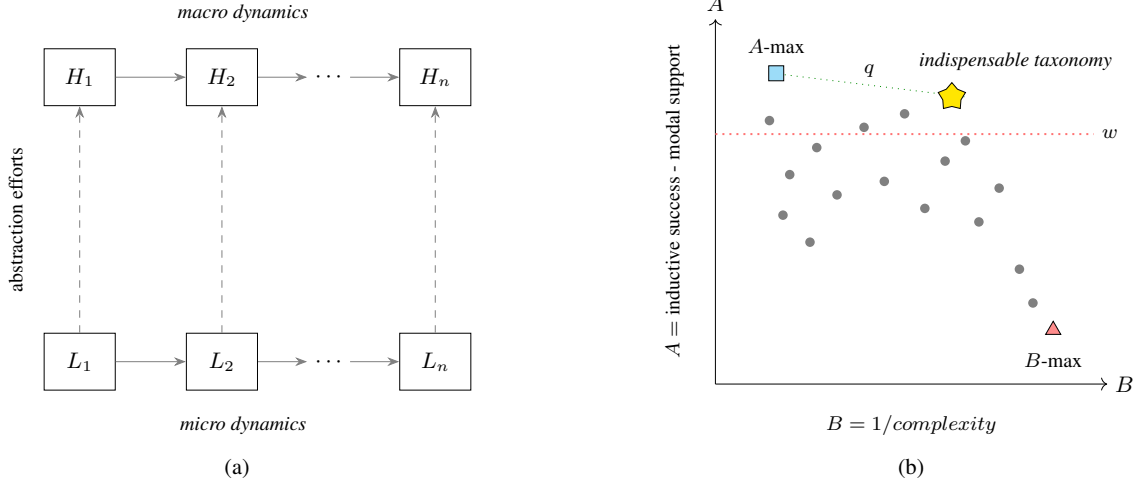


Figure 1: (a) Possible paths from microdynamics to a target macrovariable H_i of H_n , through abstraction and macro-/micro-dynamical evolution. **(b)** A task is solvable by different effective taxonomies (each dot) that generate the dynamics of generalizations with a fidelity down to some noise tolerance, here depicted by threshold w ; it is a non-trivial, empirical matter whether those taxonomies exist, and how they are distributed. In this toy example, the dots show some trade-offs between inductive success/modal support (A) and inverse complexity (B). The A -max taxonomy (square) maximizes inductive success at relatively maximal complexity costs; the B -max taxonomy (triangle) is relatively minimally complex but inductively idle. The indispensable taxonomy (star) is the effective taxonomy within threshold w that has the lowest complexity; if the example task were less demanding regarding noise tolerance, both w and, in accordance, the indispensable taxonomy, would change, and the dotted line q , connecting the indispensable taxonomy with A -max, would have a more pronounced negative slope.

terion that can choose ontologies for a given solvable task, there is another obvious source that threatens to bring back some form of multiplicity and relativism *prima facie* unwanted by the realist goals of the framework, the Global Challenge: there is a dense space from which to choose different possibly solvable tasks, each potentially yielding their own new effective taxonomies. Indispensability targets the Local Challenge for a given task, but in its current form cannot address the Global Challenge—precisely because on many occasions there might be a trade-off between the Local and the Global Challenge. Imagine that by some fluke a particle p is exactly at the centre of mass of Neptune for the next 10 years; we can solve the task of tracking the position of Neptune for the next 10 years by discovering the effective theory “follow this exact particle p ”; having an effective theory that implies calculating centres of mass is more complex for the task than simply tracking p ; this formally results in particle p as the indispensable ontology; however, having centres of mass would clearly be inductively better (think about the 11th year of trying to track Neptune) and would generalize better (p is not present in other planets). Considerations about robustness in induction and extrapolation seem to have a weight for scientific realist considerations that is not captured by indispensability alone.

Shalizi and Moore wittily comment on what we call here the Global Challenge [23, p.17]:

Our initial choice of observables is a function of our ability and inclinations to observe the system, and our choices about what to predict. Beyond that initial choice, and the requirement that good macrostates have certain causal properties, the causal states we use are completely out of our control, fixed entirely by the system’s objective, microphysical dynamics. A different set of initial observables will, generally, lead to a different set of causal states. Sometimes, but not always, these causal states are related in a hierarchy of emergence. One might put it like this: for every question we ask It, Nature has a definite answer;

but Nature has no preferred questions.

This situation gives rise to the following soft **dilemma**: eMSC approaches can **EITHER** (Horn 1) develop accounts that outline how much and what type of taxonomical multiplicity is acceptable while avoiding undermining the ontic and scientific status of effective theories, **OR** (Horn 2) try to develop a solution or approximation to the open Global Challenge.

The strategies constitute a dilemma when they are defended in their stronger versions where one excludes the other; but softer versions of both positions can be developed by conceding points to the contrary strategy. Regarding Horn 1, it is still philosophically underexplored whether the realist goals of eMSC are compatible with a form of multiplicity that would be satisfactory for the framework. This is an open question, which is being indirectly and partially explored in different but related contexts (e.g., in discussions over the compatibility of scientific realism with the abandonment of the natural/non-natural kinds distinction: [24, 6]). Regarding Horn 2, the philosophers Ladyman [12] and Ross [14] have indirectly pointed to a way to have a further scarcity criterion, something which seems to target the worries of the Global Challenge: we should aim at effective taxonomies that can be part of a systematic process of scientific unification, and coherent triangulation between the ontic commitments of different domains. This move, however, has not been clearly developed anywhere in the philosophical literature (this shortcoming has been highlighted by [3, 18]). Other intellectually related projects, such as William Wimsatt’s “Robustness Analysis” [29], are also underexplored in the sense we care about here.

After providing this framing that underlines the Global Challenge, this article asks: while its philosophical side remains only timidly explored, has the scientific side of eMSC addressed it with more success? The short answer is: no, but there are tools from the broader computational modelling field that directly fit the Challenge. The different scientific modelling proposals within the eMSC framework

try to provide objective criteria for choosing taxonomies at some scales over others by analysing the dynamics of interaction between the different possible scales of specific systems. As a result, interesting compositional properties between scales have been defined (e.g., causal emergence, effective information, synergistic information, and computational closure), and different empirical techniques have been proposed [32, 11, 20]. However, while formal developments around scales and mereology might be a crucial part of the necessary work for the eMSC project to be successful, a focus on isolated systems does not address the dilemma. The dilemma can be addressed by exploring the possibilities and limits of Horn 1, and/or by evaluating and developing concepts and tools that potentially tackle Horn 2 by exploring when and why patterns that are found in very distinct systems at different scales are modally equivalent. On this latter point, eMSC framework could integrate active discussions regarding multiple realizability [19], clustering and machine learning methods [28], as well as analyses of the limits of induction [21, 30, 25] and across-system extrapolation (see [2]; or, especially relevant for eMSC, the latest extensions of computational mechanics [4]).

In sum, the point raised by this article is that the goals of eMSC approaches may call for complementary work that currently receives insufficient attention. We might discover that there are principled reasons why the Global Challenge is unsolvable, either in general or for particular fields. This would imply that there is no scientifically principled justification for their effective taxonomies, and—importantly—that extra-theoretical, practical considerations would be needed to carry those scientific inquiries forward.

ACKNOWLEDGEMENTS

I would like to thank the BM group, and especially my supervisor Manolo Martínez, for stimulating discussions on this topic. I am also grateful to James Ladyman, David Wolpert, and Abel Jansma for their welcoming engagement with my questions. This research was made possible by a BIAP Maria de Maeztu FPI grant (CEX2021-001169-M-20-1).

REFERENCES

- [1] Holly K. Andersen, ‘The density of structure’, *Synthese*, **206**(5), 237, (2025).
- [2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan, ‘A theory of learning from different domains’, *Machine Learning*, **79**(1–2), 151–175, (2010).
- [3] Majid Davoody Beni, ‘Structural realism, metaphysical unification, and the ontology and epistemology of patterns’, *International Studies in the Philosophy of Science*, **31**(3), 285–300, (2017).
- [4] Nicolas Brodu and James P. Crutchfield, ‘Discovering causal structure with reproducing-kernel hilbert space ϵ -machines’, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **32**(2), (2022).
- [5] Sean M. Carroll and Achyuth Parola, ‘What emergence can possibly mean’, in *Dennett’s Real Patterns in Science and Nature*, eds., Tyler Millhouse, Steve Petersen, and Don Ross, MIT Press, Cambridge, MA, (2026).
- [6] Anjan Chakravartty, ‘Last chance saloons for natural kind realism’, *American Philosophical Quarterly*, **60**(1), 63–81, (2023).
- [7] Daniel C. Dennett, ‘Real patterns’, *The Journal of Philosophy*, **88**(1), 27–51, (1991).
- [8] Joe Dewhurst, ‘Causal emergence from effective information: Neither causal nor emergent?’, *Thought: A Journal of Philosophy*, **10**(3), 158–168, (2021).
- [9] Frederick Eberhardt and Lin Lin Lee, ‘Causal emergence: When distortions in a map obscure the territory’, *Philosophies*, **7**(2), (2022).
- [10] Peter Grassberger, ‘Some comments on computational mechanics, complexity measures, and all that’, 2018.
- [11] Abel Jansma and Erik Hoel, ‘Engineering emergence’, arXiv preprint, 2025.
- [12] James Ladyman, ‘Patterns all the way up: Prolegomena to a future naturalized metaphysics’, in *Dennett’s Real Patterns in Science and Nature*, eds., Tyler Millhouse, Steve Petersen, and Don Ross, chapter 3, MIT Press, Cambridge, MA, (2026).
- [13] James Ladyman and Lorenzo Lorenzetti, ‘Effective ontic structural realism’, *The British Journal for the Philosophy of Science*, (2023). Online ahead of print; forthcoming in vol. 78.
- [14] James Ladyman and Don Ross, *Every Thing Must Go: Metaphysics Naturalized*, Oxford University Press, Oxford, 2007.
- [15] James Ladyman and Karoline Wiesner, *What Is a Complex System?*, Yale University Press, New Haven, CT, 2020.
- [16] Sarah E. Marzen and Simon DeDeo, ‘The evolution of lossy compression’, *Journal of The Royal Society Interface*, **14**(130), 20170166, (2017).
- [17] Patricia Palacios, *Emergence and Reduction in Physics*, Cambridge Elements in Physics, Cambridge University Press, Cambridge, 2022.
- [18] Oriol Roca-Martín, ‘A New Problem for Real Patterns Ontic Structural Realism?—Separating Local from Global Indispensability’, *Manuscript in preparation for submission*, (2026).
- [19] Oriol Roca-Martín and Manolo Martínez, ‘What Can Multiple Realizability Possibly Be? A Complexity Approach’, *Manuscript in preparation for submission*, (2026).
- [20] Fernando E. Rosas, Bernhard C. Geiger, Andrea I. Luppi, Anil K. Seth, Daniel Polani, Michael Gastpar, and Pedro A. M. Mediano, ‘Software in the natural world: A computational approach to hierarchical emergence’, arXiv preprint, 2024.
- [21] Gerhard Schurz, *Hume’s Problem Solved: The Optimality of Meta-Induction*, MIT Press, Cambridge, MA, 2019.
- [22] Cosma Rohilla Shalizi and James P. Crutchfield, ‘Computational mechanics: Pattern and prediction, structure and simplicity’, *Journal of Statistical Physics*, **104**(3–4), 817–879, (2001).
- [23] Cosma Rohilla Shalizi and Christopher Moore, ‘What is a macrostate? subjective observations and objective dynamics’, *Foundations of Physics*, **55**(1), 2, (2025).
- [24] Matthew H. Slater, ‘Less work for theories of natural kinds’, *The Philosophical Quarterly*, (2024).
- [25] Tom F. Sterkenburg and Peter D. Grünwald, ‘The no-free-lunch theorems of supervised learning’, *Synthese*, **199**(3–4), 9979–10015, (2021).
- [26] Abel Suñé and Manolo Martínez, ‘Real patterns and indispensability’, *Synthese*, **198**(5), 4315–4330, (2021).
- [27] Thomas F. Varley, ‘Information theory for complex systems scientists: What, why, and how’, *Physics Reports*, **1148**, 1–55, (2025).
- [28] David S. Watson, ‘On the philosophy of unsupervised learning’, *Philosophy & Technology*, **36**(2), 28, (2023).
- [29] William C. Wimsatt, ‘Robustness, reliability, and overdetermination’, in *Scientific Inquiry and the Social Sciences*, eds., Marilynn B. Brewer and Barry E. Collins, 124–163, Jossey-Bass, San Francisco, (1981).
- [30] David H. Wolpert, ‘The implications of the no-free-lunch theorems for meta-induction’, *Journal for General Philosophy of Science*, **54**(3), 421–432, (2023).
- [31] David H. Wolpert, Joshua A. Grochow, Eric Libby, and Simon DeDeo, ‘Optimal high-level descriptions of dynamical systems’, *arXiv e-prints*, arXiv-1409, (2014).
- [32] Bing Yuan, Zhang Jiang, Aobo Lyu, Jiayun Wu, Zhipeng Wang, Mingzhe Yang, Kaiwei Liu, Muyun Mou, and Peng Cui, ‘Emergence and causality in complex systems: A survey on causal emergence and related quantitative studies’, arXiv preprint, 2024.
- [33] J. Zhang, R. Tao, K. H. Leong, et al., ‘Dynamical reversibility and a new theory of causal emergence based on svd’, *npj Complexity*, **2**, 3, (2025).

Concrete computations and virtual machines: In defence of a computational functionalism

Simon Bowes ¹

Abstract. Digital computation may not be sufficient for consciousness, but this doesn't warrant biological naturalism. Rejecting coarse-grained Turing-computational functionalism and adopting a broader computationalism with no built-in syntax/semantics distinction, undermines the motivation for biological naturalism. Substrate dependent processes of the kind found in living things may be necessary for the emergence of consciousness, but these are functional requirements, not intrinsic properties.

1 Main text

Seth (2026) argues against computational functionalism to rule out the possibility of consciousness arising in computing machines. I agree that to be conscious, "artefacts may have to be closer to biological systems than allowed for by conventional digital computation"², but not that this implies Biological Naturalism: 'closer' can be functionally defined using 'unconventional' computation. Intelligence and consciousness partly dissociate, but not because one is functional the other intrinsic. Unconscious algorithms produce intelligent behaviour, but if consciousness evolved, it is because it does too. If it additionally has intrinsic, non-functional properties, Zombies lurch into view.

Consciousness is essentially about what a system does – we are conscious for a reason. 'Feeling like something' doesn't imply intrinsicity. Experiences have the character they do due to 'bottom-up' and 'top-down' processes; to form unified experiences with intrinsic properties, these would need to be added together somehow. This is a form of Cartesian Materialism, which there are reasons to avoid (Dennett, 1991). However, my focus here will be the supposed incompatibility of computation and consciousness.

It's true that being usefully describable in terms of computations does not entail the brain computes, just as we might usefully describe a falling brick computationally, without assuming the brick computes. But we have independent reasons to say the brain computes, namely, intentionality. Mental states are about things, they are informational, and the processes they are involved in take this information and use it to control behaviour, and so they can be said to compute, in a way bricks don't.

'Coarse-grained' functionalism is too liberal, in that something could satisfy a functional description of a mental process without thereby being a mental process, but finer-grained, more demanding varieties of functionalism are available. Turing-computational functionalism is particularly coarse-grained, but '4E' cognitive science supports other kinds of computationalism. Virtual Machine Func-

tionalism (Sloman & Chrisley, 2003; Bowes, 2023), captures 'generatively entrenched' organisational features required for mindedness, avoiding overly liberal functionalism: parts play their roles due to (contextual, dispositional) causal powers. The "limited substrate flexibility" of neurobiological functions makes Turing-computational functionalism unsuitable, but there are "broader forms of computation which have similar constraints." Contra Seth, broadening our understanding of computation doesn't leave computational functionalism without unmotivated, but undermines the argument for biological naturalism.

Embodied and embedded forms of computation are not "pale shadows" of life. There may be an "excitable fluidity" to biological entities ripe for the emergence of consciousness, but why think only biological entities can have this? The computers we build are 'rigid' because we design them to blindly follow rules to match the semantics we assign to symbols. But computation doesn't necessitate a strict syntax/semantics separation:

"... a fundamental theory of programs, and of computation, should acknowledge that an essential function of a computer program is to make things happen. Whereas symbolic logic can be viewed as mere playing around with uninterpreted formal calculi..., and computational logic can be seen as the study of abstract timeless relations in mathematically specified 'machines' (such as Turing machines), computer science cannot properly be described in either of these ways. It follows... that the familiar characterization of computer programs as all syntax and no semantics is mistaken." (Boden, 1988a)

The point is that arguments like Searle's (1980) depend on the artificial separation of implementation from semantics, whereas in embodied and embedded computational systems, these cannot be separated.

There are two ways to interpret the claim that (Turing) computation is not sufficient for consciousness: being describable in terms of computations is not enough, and being computational is not enough.

A computational description is given in terms of algorithms we can specify, but these are models of systems, and to think that models implement what they model is to conflate a map with the territory. We build systems to implement these models, but shouldn't expect real-world systems to operate on the same principles. We build them, imposing meanings on their operations, but without us they don't have any semantics. As such, when building computers we use the algorithms as engineering blueprints rather than scientific models. We shouldn't conflate engineering uses of computations with scientific uses of computational concepts (Boden, 1988b, Rooij, 2024).

When modelling rather than making, we can be using the model loosely or tightly. In a loose use, the model is not taken to reveal

¹ University of Sussex, England.
s.c.bowes@sussex.ac.uk

² Unattributed quotes are from Seth (2026).

structural facts about the system, but is an as if use. We can treat the tree as if it's calculating the optimal shape to grow into the prevailing winds. It isn't in fact doing any calculating, but the strategy allows some useful inferences. In a tight use, the inferential usefulness of the model is explained by an empirical relationship between the structure of the model and the structure of the system. A computational description may be useful if the system being modelled computes, in that it uses information to inform goal-directed behaviour. This does not necessarily mean the system is manipulating symbolic representations, but may be computational in one of the broader senses.

Brian Cantwell Smith (1996) shows how some of the standard theories of computation are inadequate. Four such theories are: 1 – Computation is formal symbol manipulation 2 – Effective computability – Turing machines 3 – Computing is information processing 4 – Computers are digital systems To dismiss them in order: formal symbol systems are described logically not mechanically, subject to the non-local semantic norms implicit in the logical system constructed by the designers, so this cannot be an adequate description of a concrete computational system. Effective computability doesn't deal with meaning at all, it's pure maths. Again, computer scientists take semantics and 'dump it in the mechanism'. Information is a semantic notion, not a physical quantity like heat. Computers are causally engaged in the actual world, so an account of how a concrete system comes to be about things in the world in which it operates is needed. Digitality is just an engineering notion, a way of dealing with the messiness of the world, rather than a theory of computation.

Aaron Sloman (2002) agrees that narrow, syntactic, formal definitions of computation (manipulations of digital symbols by Turing machines) are inadequate for explaining how the mind works, and that computation should be defined in a much broader, ontological and causal framework: computation is the operation of a control system driven by information processing.

This view of computation is distinguished from traditional computationalism in remembering that real-world computational systems need to operate in the real world. Levels of explanation (e.g. Marr's 3 levels) may be misinterpreted as fully independent realms, rather than mutually constraining levels of the organisation of stuff. Some causal claims at higher explanatory levels can be true, abstraction is permitted, but there is always a debt accrued, which cannot be left unsettled indefinitely. Epistemological prudence, then, requires an account of the computational level that keeps an eye on things. This is provided by Virtual Machine Functionalism.

Virtual Machine Functionalism (VMF) argues that human mentality (consciousness, thoughts, emotions) does not happen at the physical "hardware" level (e.g. of neurons), but at a "virtual machine" level. This is not a simulated machine, but is a coalition of parts brought together in real time to form a structure that works towards a goal. It is designed to be less tied to the physical level than a machine of discrete parts with definite functions (like a clockwork mechanism), but not completely untethered from the material substrate as algorithmic descriptions. A computational system is a web of complex, concurrent, interacting virtual states. E.g. The events in a chess program are real virtual events because they have causal consequences—even though they cannot be defined solely by physical laws.

Computation is not just crunching data in bits; it is the dynamics of information change used in system control, to govern action, navigate an environment, evaluate goals, and make choices. A machine computes to stay in control of its survival or its tasks (Sloman, 1992). This is in opposition views of computation as all syntax. Mathematical definitions of computation are static, structural, and formal. A

static string of mathematical equations on a piece of paper satisfies the formal definition of an algorithm, but is not a mind. A mind requires active, running processes that interact causally with one another and the environment (and given the functional requirements for operating in our world, this will restrict the possible instantiations). Traditional computation is discrete. Real mentality requires continuous (chemical, biological) mechanisms.

Both Sloman and Smith agree that Turing machines, formal syntax, and digitality fail to explain what computers actually do in the wild, or how the human mind operates. Smith redefines computation as part of a theory of meaning – how matter and mechanisms can be about things. Sloman redefines computation as an engineering framework of informational control.

The inherently semantic nature of computation informs Boden's English Reply to Searle's Chinese Room challenge to computational functionalism (Boden, 1988a, 260-64):

The crux of the English reply is that the instantiation of a computer program, whether by man or by manufactured machine, does involve understanding - at least of the rule-book. Searle's initial example depends critically on Searle-in-the-room's being able to understand the language in which the rules are written, namely English. . . . B. C. Smith (1982) argues that programmed representations, too, are inherently active-and that an adequate theory of the semantics of programming-languages would recognize the fact. . . . computer scientists commonly make too complete a theoretical separation between a program's control-functions and its nature as a formal-syntactic system. . . . it would be more elegant and less confusing to adopt a 'unified' theory of programming-languages, designed to cover both denotative and procedural aspects. . . . It follows from Smith's argument that the familiar characterization of computer programs as all syntax and no semantics is mistaken. The inherent procedural consequences of any computer program give it a toehold in semantics, where the semantics in question is not denotational, but causal. The analogy is with Searle-in-the-room's understanding of English, not his understanding of Chinese.

So, natural minds are embodied and embedded physical computers, with representational content derived from engagement with a world, and substrate dependence that enables "relevant patterns of functional organisation."

Seth uses the lack of substrate flexibility against embodied computation as it "might not be implementable in. . . non-biological substrates." But, if radical substrate independence makes Turing computation unsuitable for consciousness, the need to be embodied and embedded in the right way should be seen an advantage for a broad-minded computationalism. There might be other, non-biological substrates that minds could be instantiated in – if minds only exist in biological organisms, this would be a contingent rather than necessary fact.

The way to avoid overly liberal functionalism is to foreground embodiment: momentarily matching an abstract description is insufficient to realise a computation. There needs to be some substrate flexibility though, so that relevantly similar mental states can be realised in reasonably varied embodiments. Maybe only biological substrates currently implement the right kind of computation, but knowing what makes them suitable should make replication possible. Can we just stipulate that these 'replicants' are biological, thus making biological naturalism true by fiat? We could, but whether we expand or restrict such 'natural' categories is only partly determined by the way things

are, also depending on the usefulness of the categories given our explanatory interests (Bowes, 2023).

Accepting that ‘excitable’ autopoietic processes are responsible for the emergence of life as we know it, then such processes were sufficient for the emergence of consciousness. But were they necessary, and would that support biological naturalism? The claim that consciousness came about through natural processes, and that carbon-based life on earth is one way this could happen, is just naturalism.

‘Life’ may be defined by “homeostatic processes... linked to higher-level functional properties,” which are emergent, in that “knowledge of (or interventions at) lower, microscopic levels does not add anything to macro-level predictability, or help determine macro-level outcomes”. Seth says consciousness could depend on functional or intrinsic properties of the biological substrate, but that, given emergence, it is the properties at the level of life that matter. But the dependence is stronger, otherwise the higher-level properties would be inexplicable – strongly emergent in the sense of ‘popping out’ of nowhere.

Seth claims “apparently intrinsic properties at one level might always be understandable in terms of patterns of functional organisation at more fine-grained levels.” This would make intrinsicity perspectival, related to our explanatory interests, but conflates intrinsicity with explanatory autonomy. Instead of being ‘intrinsic,’ non-relational properties, they are irreducible – emergent, but not in the sense of inexplicably appearing. To be explicable, and so of scientific interest, they cannot be intrinsic, as explanation requires relations (just not necessarily reductive ones). This implies inherently irreducible, rather than intrinsic, causal powers.

The higher-level causal properties of mindedness, then, are computational in a broad sense; being biological is part of the “finer-grained functional substrate” out of which they may emerge. The additional claim of necessity depends on an assumption that only biological entities can have the suitable properties, which has not been established. This leaves the door open to a form of causally constrained computational functionalism.

Seth’s argument seems to be based on a Cartesian intuition regarding the intrinsicity of consciousness (that it’s not “about what a system does”), against functionalist accounts. This is in tension with his ‘real problem’ approach, and is prone to zombie attack: given the substrate’s ‘finer-grained’ properties, could there still be something missing, some intrinsic property that emerges in a way that isn’t explained by those functional facts plus their embodied history? If not, we should accept functionalism. If so, then the hard problem ‘pops up’ again.

2 REFERENCES

- Boden, M. A. (1988a). Escaping from the Chinese room. In *Computer Models On Mind: Computational Approaches In Theoretical Psychology*. Cambridge University Press.
- Boden, M. A. (1988b). *Computer Models On Mind: Computational Approaches In Theoretical Psychology*. Cambridge University Press.
- Bowes, S. (2023). *Naturally Minded: Mental Causation, Virtual Machines, and Maps*. Springer Verlag.
- Dennett, D. C. (1991). *Consciousness Explained*. Penguin Books.
- Rooij, Iris & Guest, Olivia & Adolfs, Federico & Haan, Ronald & Kolokolova, Antonina & Rich, Patricia. (2024). Reclaiming AI as a Theoretical Tool for Cognitive Science. *Computational Brain & Behavior*. 7. 10.
- Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*(3), 417–57.

Seth AK. (2025) Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*: 1-42

Sloman A. (1993) The Mind as a Control System. *Royal Institute of Philosophy Supplement*. 34:69-110. doi:10.1017/S1358246100002460

Sloman, A.(2002). The irrelevance of Turing machines to artificial intelligence. In Matthias Scheutz, *Computationalism: New Directions*. MIT Press.

Sloman, A. & Chrisley, R. (2003). Virtual Machines and Consciousness. *Journal of Consciousness Studies*. 10. 133-172.

Smith, Brian Cantwell (1996), *On the Origin of Objects* (Cambridge, MA: MIT Press)

Even if humans can answer questions Turing machines cannot, that doesn't mean that digital computational consciousness is impossible

Ron Chrisley ¹

Diagonal arguments against algorithmic accounts of consciousness date back to Lucas, Turing and Gödel. Penrose's formulation, which is the focus of this talk, purports to establish not only the impossibility of algorithmic consciousness, but the impossibility of a digital computer to even simulate the behaviour of, e.g., conscious human mathematicians. To this end, Penrose constructs a question, which we can call the TM Self-Halting Question ("does the k th Turing machine not halt on input k ?") and demonstrates that its characteristic function is not Turing computable. Further, that we can understand his argument demonstrates that we, by contrast, can answer the TM Self-Halting Question for arbitrary k (because we know that for all k , if TM_k is sound with respect to the TM Self-Halting Question, then it does not halt on input k).

Building on an insight from Whiteley, I show that even if one grants Penrose the above results, his negative conclusions concerning the inability of computational systems to simulate consciousness do not follow. This is a consequence of his largely unexamined assumptions concerning simulation. In particular, I show that x can perfectly simulate y even if x and y do not compute the same class of functions; I achieve this in four steps:

1. Characterizing the class of (Turing) computational systems as members of a more general class of question-answering systems.
2. Showing that for every class of question-answering systems (even finite ones) there exists a question relatively isomorphic to the original TM Self-Halting Question.
3. Showing how two question-answering systems may compute different functions, and yet compute the same class of function types.
4. And arguing that for x to perfectly simulate y , x need only compute the same class of function types as y .

Separately, and more briefly, I refute Penrose by arguing that x can be conscious even if there exist conscious agents y that x cannot simulate.

These results each disarm the diagonal argument for the non-algorithmic nature of human conscious behavior, and thus make room for a classical, digital, computational replication and understanding of consciousness.

REFERENCES

[1] R. Penrose, 'Shadows of the mind: A search for the missing science of consciousness'. Oxford University Press. (1994).

[2] Whiteley, C. H., 'Minds, Machines and Godel: A Reply to Mr Lucas.' *Philosophy* 37(139):61-62. doi:10.1017/S0031819100030886 (1962).

¹ University of Sussex, England,
R.L.Chrisley@sussex.ac.uk

What Computation Explains, and What It Does Not about Daily Cognitive Activity?

Gutoreva, A.* and Omar, Z. and Bektemisov, N. and Nurlankyzy ¹

Abstract.

We present a multimodal, free-living study of daily cognitive activity combining frontal EEG (Neurable MW75), Apple Watch SE biometrics, and ecological momentary assessment across two weeks with eight participants, yielding 1,904 annotated ten-minute segments across nine activity domains. Fusing these heterogeneous streams, we find that motivation and attentional focus track circadian structure while energy does not, that Focus–Motivation coupling ($r_s = +0.71$) anchors a structured cognitive activation network, and that a BiLSTM with cross-modal attention achieves flow state classification AUC=0.758 under leave-one-subject-out cross-validation. We situate these results at Marr’s algorithmic level [1], and examine three specific ways in which the computational framework cannot recover the phenomenological dimension of cognitive daily activity.

1 Introduction

What are the benefits and limits of treating cognition computationally? This paper offers an empirical answer grounded in a concrete case: a multimodal, free-living study of daily cognitive activity that deploys computational methods at the boundaries of their explanatory reach, and examines what, if anything, they leave behind.

We recruited eight participants to wear an EEG headset (Neurable MW75), an Apple Watch SE, and to complete ecological momentary assessment (EMA) [2] surveys across two weeks of naturalistic daily life. Fixed ten-minute intervals produced 1,904 annotated segments spanning nine activity domains, yielding self-reported ratings of motivation, attentional focus, arousal, energy, perceived stress, emotional valence, and flow state. These heterogeneous streams — neural, physiological, behavioural, and phenomenological — were fused through shared temporal epoch alignment and analysed using OLS regression, Kruskal–Wallis tests, Spearman correlations, and a Bidirectional Long Short-Term Memory (BiLSTM) model with four-head cross-modal attention.

The pipeline processes three parallel streams — EMA self-reports, frontal EEG, and Apple Watch SE biometrics — through four sequential stages of preprocessing before merging them via cross-modal attention into a shared 128-dimensional representation evaluated under Leave-One-Subject-Out (LOSO) cross-validation.

We situate this framework at Marr’s algorithmic level [1], specifying what transformations are performed over representations, without claiming that EEG band power ratios or HRV indices constitute cognitive computation at the implementational level — a distinction that bears directly on interpreting the limits of our findings.

2 What Is Gained

Diurnal structure. Treating motivation and attentional focus as computational targets, Kruskal–Wallis tests confirm that both differ significantly between morning and afternoon periods after Bonferroni correction (motivation: $H = 11.51$, $p < .001$; focus: $H = 10.35$, $p = .001$; adjusted threshold $\alpha/3 = .017$; $N = 7$, one participant excluded due to insufficient segments). Energy does not differ significantly across the day ($H = 0.03$, $p = .87$), a finding consistent with systematic reviews showing that the type of cognitive function — not arousal generally — determines whether time-of-day effects emerge [3]. Motivation and attentional engagement, not raw energy, are the primary cognitive variables tracking circadian structure in naturalistic student populations.

Cognitive activation network. Treating cognitive dynamics as a relational network, Spearman correlations reveal a structured activation system: Focus–Motivation coupling is the strongest link ($r_s = +0.71$, $p < .001$), followed by Motivation–Energy ($r_s = +0.59$) and Focus–Energy ($r_s = +0.49$), while Energy–Stress shows a reliable negative relationship ($r_s = -0.42$, $p < .001$). These associations constitute an empirically derived map of how cognitive resources co-vary in free-living conditions, providing a quantitative foundation for intervention design.

Temporal flow classification. The BiLSTM model with cross-modal attention achieves flow state classification AUC=0.758 and F1=0.575 under LOSO-CV across six test folds, exceeding the biometrics-only baseline by 61.5% and the early-fusion Random Forest baseline by 6.5%. A 2-layer BiLSTM with hidden size 64 processes sliding windows of six consecutive segments (60 minutes), encoding bidirectional context so that each 10-minute epoch is interpreted in light of both its preceding and subsequent trajectory. This naturalistic full-day approach addresses a gap identified by recent work: Knierim et al. [4] show that flow during real-world knowledge work can be detected with wearable EEG across an entire working day, while Rácz et al. [5] confirm that combined EEG, HR, and SpO₂ signals from consumer devices reliably distinguish flow from anti-flow states. Recent multimodal work further confirms the value of EEG–physiological fusion: Salam et al. [6] demonstrate that combining EEG and ECG features improves cognitive stress classification to 85.7% accuracy in *Scientific Reports*, underscoring that cross-modal fusion consistently outperforms single-modality approaches. We plan to evaluate Transformer-based temporal fusion and within-subject meta-learning against the BiLSTM baseline once the full cohort ($n = 20$) is collected, with SHAP-based attribution to resolve the EMA-versus-fusion anomaly.

¹ School of Information Technology and Engineering
Kazakh-British Technical University, Almaty, Kazakhstan
* a.gutoreva@kbtu.kz

3 What Is Lost

The computational account predicts the occurrence of flow with reasonable accuracy, but it cannot recover three things that matter for a complete account of cognition.

First, it cannot distinguish between phenomenologically distinct experiences that share the same multimodal signature. Two participants in this dataset show overlapping EEG theta/beta ratios and comparable HRV profiles during structurally similar tasks, yet report qualitatively different subjective experiences. This is not a calibration artefact. Cuve et al. [7] demonstrate, using a multivariate autonomic-subjective affective space across 51 participants, that substantial autonomic structure exists that simply does not covary with subjective experience — a second dimension of physiological variance orthogonal to self-report, invisible to any model trained on labels alone. The phenomenological level — Marr’s third level — is not represented in the feature space, and no amount of additional sensor modalities can recover it.

Second, the framework cannot capture the retrospective reconstitution of experience. Participants in naturalistic EMA studies frequently revise in-the-moment ratings when reflecting in aggregate, indicating that phenomenological meaning is not a point-in-time measurement but an interpretive act distributed across time. Leertouwer et al. [8] show that retrospective affect summaries systematically diverge from aggregated EMA means in ways predicted by individual interpretive strategies rather than by the underlying moment-to-moment data — confirming this is not a data quality problem solvable by increasing sampling resolution but a structural feature of how experience is organised.

Third, and most fundamentally, flow as theorised by Csikszentmihalyi [9] is not merely correlated with certain biometric conditions — it is partly constituted by the subjective quality of the experience itself. A model that classifies flow from EEG spectral features and HRV indices is predicting the *conditions* under which flow tends to arise, not flow as such. Hang et al. [10], using single-channel prefrontal EEG across six multifaceted flow-induction tasks in naturalistic settings, find that EEG correlates explain at most 16.3% of variance in subjective flow scores — leaving the large majority of the phenomenological experience unaccounted for. This is sharpened by Akiyama et al. [11], who find no correlation between frontal-midline theta and subjective flow across more than 700 gameplay sessions in six of seven participants — directly challenging the assumption that canonical EEG signatures reliably index the experience. The gap between predictive adequacy and constitutive explanation is real, and collapsing it would misrepresent what the framework achieves.

4 Hypotheses for the Full Study

H1: Morning–afternoon differences in motivation and attentional focus will replicate across the expanded cohort while energy remains non-significant, reflecting the asymmetric relationship between circadian structure and different cognitive resource types.

H2: The BiLSTM fusion model will surpass the $F1 \geq 0.70$ threshold with expanded training data and EEG-based flow label validation via the Neurale SDK API, resolving the current partial support driven by extreme inter-individual variability in self-reported flow prevalence (range: 13–95% across the eight-participant cohort).

H3: The cognitive activation network structure — particularly the Focus–Motivation coupling — will remain stable across the expanded cohort, confirming it as a robust feature of naturalistic student cognition rather than an artefact of the current small sample.

H4: Self-reported flow labels will show systematic divergence from

EEG-based flow signatures in a non-trivial proportion of segments, and this divergence will co-vary with activity type and time of day in theoretically interpretable ways — specifically, divergence will be highest during routine activities in the late afternoon (flow-like absorption without neural signatures of optimal engagement) and lowest during purposeful morning work sessions where phenomenological and implementational indicators converge. If confirmed, H4 would constitute direct empirical evidence that the phenomenological and computational dimensions of flow are partially dissociable not merely because measurement is imperfect, but because the constructs themselves are not identical.

5 Conclusion

The computational treatment of cognition purchases genuine explanatory traction at the algorithmic level. Motivation, attention, flow probability, and their temporal interdependencies can be modelled, predicted, and mapped with consumer-grade hardware in fully naturalistic conditions — removing the laboratory as a prerequisite for cognitive monitoring research and establishing a reusable annotated benchmark of 1,904 segments for future multimodal studies.

What the framework does not purchase is an account of the phenomenological dimension of cognitive life. The specific ways in which that gap manifests — signature equivalence without experiential equivalence, retrospective reconstitution of meaning, and the constitutive rather than merely correlational character of flow — are themselves theoretical contributions. Acknowledging that boundary clearly, and building testable hypotheses around it, is part of what distinguishes a mature computational cognitive science from one that mistakes predictive adequacy for explanatory completeness.

Keywords: multimodal cognitive monitoring; ecological momentary assessment; diurnal variation; Marr’s levels of analysis; flow state; computational explanation; naturalistic cognition

6 References

- [1] Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman.
- [2] Larson, R., & Csikszentmihalyi, M. (1983). The experience sampling method. *New Directions for Methodology of Social and Behavioral Science*, 15, 41–56.
- [3] Munnilari, M., Bommasamudram, T., Easow, J., Tod, D., Varamenti, E., Edwards, B. J., et al. (2023). Diurnal variation in variables related to cognitive performance: a systematic review. *Sleep and Breathing*, 28, 611–626. <https://doi.org/10.1007/s11325-023-02895-0>
- [4] Knierim, M. T., Stano, F., Kurz, F., Heusch, A., & Wilson, M. L. (2025). Exploring flow in real-world knowledge work using discreet cEEGrid sensors. *arXiv preprint arXiv:2501.19217*. <https://doi.org/10.48550/arXiv.2501.19217>
- [5] Rácz, M., Becske, M., Magyaródi, T., Kitta, G., Szurómi, M., & Márton, G. (2025). Physiological assessment of the psychological flow state using wearable devices. *Scientific Reports*, 15, 11839. <https://doi.org/10.1038/s41598-025-95647-x>
- [6] Salam, A., Alam, F., Shah, D., Rahman, S. U., Ali, S., & Tahir, M. (2026). Improving cognitive stress classification via multimodal EEG and ECG fusion: gender differences in physiological response. *Scientific Reports*, 16, 7304. <https://doi.org/10.1038/s41598-026-38356-3>
- [7] Cuve, H. C., Harper, J., Catmur, C., & Bird, G. (2023). Coherence and divergence in autonomic-subjective affective space. *Psychophysiology*, 60(6), e14262. <https://doi.org/10.1111/psyp.14262>
- [8] Leertouwer, I. J., Schuurman, N. K., & Vermunt, J. K. (2022). Are retrospective assessments means of people’s experiences? *Journal for Person-Oriented Research*, 8(2), 52–70. <https://doi.org/10.17505/jpor.2022.24855>

- [9] Csikszentmihalyi, M. (1990). *Flow: The Psychology of Optimal Experience*. Harper & Row.
- [10] Hang, Y., Unenbat, B., Tang, S., Wang, F., Lin, B., & Zhang, D. (2024). Exploring the neural correlates of flow experience with multifaceted tasks and a single-channel prefrontal EEG recording. *Sensors*, 24(6), 1894. <https://doi.org/10.3390/s24061894>
- [11] Akiyama, T., et al. (2024). The flow state is not accompanied by frontal-midline theta activity: An EEG investigation of more than 700 video gameplay sessions. *bioRxiv*. <https://doi.org/10.1101/2024.07.11.603158>

Computational Expressivity is the key to resolving Searle’s Chinese Room argument

Dean Petters¹ and Achim Jung²

Abstract. Differences in expressivity are more important than differences in Turing computability or computational complexity when considering phenomena from Cognitive Science such as intentionality, meaning and consciousness. This means that the Chinese Room Argument as it stands doesn’t work because it has nothing to say about expressivity differences in different programs or different computational formalisms. In addition, whatever scenarios Searle comes up with in the CRA that clearly demonstrate a lack in intentionality cannot generalise to all possible computational formalisms. As there is no theorem that acts as an expressivity-analogue of the Church-Turing Thesis

1 Overview

Searle’s [4] Chinese Room Argument gedankenexperiment has a person inside a room (“Searle-in-the-room”) who does not understand Chinese and who follows English instructions to process Chinese symbols according to rules without understanding the meaning of the Chinese symbols. Unbeknownst to Searle-in-the-room the outcome of his symbol processing is a pattern of actions which appear to be appropriately meaningful responses to questions posed to the Chinese room from the outside. For Searle, the persuasiveness of the CRA is that even a human executing program instructions would not be aware of the meaning of the operations they carry out inside the Chinese Room. According to Searle the CRA demonstrates that understanding cannot arise from just the operation of formal syntactic processes as operations in the Chinese room are “*syntax but not semantics*” ([4], p. 422). Therefore, in Searle’s view, the Chinese Room could not be self-aware nor ever have understanding. It is Searle’s claim of generalisation to all programs that is a critical focus of the present paper. Searle’s 1992 treatment of the CRA indicates this approach to generalisation relies on the Church-Turing Thesis (CTT):

“For our purposes, the Church-Turing thesis states that for any algorithm there is some Turing machine that can implement that algorithm. Turing’s thesis says that there is a universal Turing machine that can simulate any Turing machine. Now if we put these two together, we have the result that a universal Turing machine can implement any algorithm whatever.” ([5], p. 202)

Since the CTT refers to computability not expressivity, Searle’s original CRA scenario [4] and new variation CRA scenarios can have different formal computational properties in expressive power. We

show, in terms of making generalisations about computational expressive power, there are two types of CRA Scenario: those within and those beyond the expressivity scope of a TM. Table 1 presents eight kinds of possible variation in the kinds of computational process that might be incorporated in CRA scenarios. These variations to the CRA can lead computational formalisms to be outside the expressivity-scope of a TM. TMs and other computational formalisms with the same expressivity-scope profile can be viewed as solely concerned with situations where programs act as implementations of mathematical functions with inputs fully provided at the start of the computation and with no possibility of new inputs being included during the run of the program. The original CRA has a single locus of control (the single series of actions that Searle carries out). In addition, there is no intensional computation or self-reference in the CRA. This matters because the CTT is invoked to generalise the lessons from the CRA to *all forms of computation whatsoever*. In contrast, this paper shows that it is only legitimate to draw conclusions from the original CRA for programs and computational mechanisms which follow the basic finite input-output serial processing paradigm. If programs presented in new variants of the CRA scenario fall outside the expressivity-scope of a TM then the lessons from the original CRA will not necessarily apply. As to make generalisations from simple systems such as Searle’s original scenario to all possible systems is an unwarranted extreme reductionism. Instead, generalisations from simple systems need to be validated on a case-by-case basis for prospective program formalisms outside the expressivity-scope of the a TM. Leaving open the possibility that analysis of some other program formalism could lead to different conclusions.

The key point we make is that whilst computational formalisms do not differ in terms of Turing computability, they do differ in computational complexity and computational expressive power (expressivity/expressiveness). Our argument is that differences due to computational expressive power are critical for understanding the CRA. Differences due to complexity are limited as TMs and alternatives that are beyond the expressivity-scope of a TM compute within polynomial time of each other. We argue that expressivity is more relevant to considerations of intentionality, meaning, and understanding in computation than complexity and computability are.

The expressive power of a computational formalism is the processes and structures that can be represented and communicated in that formalism [2]. So it pertains to what kind of knowledge can be expressed in a given computational formalism. Whilst all computational formalisms are bound by Turing computability, more expressive formalisms can represent and communicate objects, entities, processes and relations that less expressive formalisms cannot ([7] p. 3). So whilst they still can only compute the same set of functions, more expressive formalisms do possess additional capabilities (that

¹ Sheffield Hallam University, UK, email: d.petters@shu.ac.uk

² University of Birmingham, UK, email: A.Jung@cs.bham.ac.uk

Domain of variation to TM processing with more expressive power than a TM	Does it change Turing computability?	Description of variations to TM processing with more expressive power than a TM
Interaction	No	ongoing interaction with the outside world
Concurrency	No	the process computational view; parallel and concurrent computation, including asynchronous parallel computations with independent speeds
Continuity	No	the kinds of ‘forever’ processes that are found in operating systems
Real time	No	more detail in how the computational system exists in time and the real world
Richer input	No*	richer forms of input, real world sensors, real number computation, continuous input, and input with infinite inputs (streams)
Indeterminancy	No	true random numbers
Self-reflection	No	routines within a system analysing their own internal processing for ‘meaningful’ patterns
Intensional computation	No	analysis of the codes of programs with code viewed as data. Including in a higher-order computational setting. Which means for programs that take programs as input, analysing these programs-as-input as data

Table 1. Eight ways in which TM processing can be varied to increase computational expressive power. These variations are all important for Cognitive Science models. Whilst none of the 8 variations to processing changes Turing computability, each of these 8 variations can change the expressive power of computational systems (*we do not claim that richer input leads to super Turing computation. However, others have made this claim [6, 1]).

may be relevant to Cognitive Science phenomena such as understanding). Representation, and the capability of one state, entity, property, or state of affairs to ‘stand for’ another is the core of intentionality. Therefore, expressivity rather than computability or complexity is more centrally relevant to what a computational formalism can represent or ‘stand for’.

To be clear about what we mean by computation within the expressivity-scope of a TM – we are referring to computation of the type that can be carried out as a ‘batch process’ where input is provided at the start (historically on cards) and output is returned on a printer. Since ‘batch computation’ refers to a technological aspect of this style of computation it is not an ideal term to use to describe this class of process. We might also use the terms: function-based; mathematical view; classic; closed; static; finite input and finite output; serial (in the sense of non-concurrent); and non-interactive (in sense of not parallel and no interruptions). From these alternatives, here we choose to refer to computation within the expressivity-scope of a TM as ‘input-output computation’ as this is a neologism and avoids confusions that might occur with some of the other possible labels set out above. We also emphasise that in referring to computation that is not input-output computation we are not referring to what is known as ‘hypercomputation’³.

We show that the variations in table 1 lead to more expressive power *and* are more credibly linked with cognitive science models and information processing in natural systems than the less expressive processing of a TM. For instance, networks of interacting persistent agents that are more intimately integrated to real world sensors, with a richer abstraction of time and space, and which are capable of self reflection, are more expressive than TMs ([7], p.3). They have greater capabilities related to computational expressive power than TMs due to what they can represent and communicate even if they can compute just the same set of computable functions as TMs. An example from the analysis of expressive power in process calculi used to describe concurrent and distributed systems is that synchronization mechanisms, such as communication and coordination, provide extra capabilities in modeling complex interactions and be-

haviours between processes [3].

We do not show how the variations in table 1 give rise to understanding and don’t need to show this to defeat the CRA. The CRA doesn’t work because the variations in table 1 can lead to programs being beyond the expressivity-scope of a TM. We cannot make generalisations about the expressivity of formalisms that produce these programs.

We highlight the observation that CRA scenarios created with computational formalisms beyond input-output computation are not only more similar to typical human cognition than the very simplified portrayal of processing in the original CRA, but the complexity they present is fast being achieved and overtaken in contemporary computing systems. Situations in contemporary computing are now so rich, they can no longer be said to be covered by a paradigm where the inputs are known in advance, the system is left alone to do its computation, and then provides the answer. Critically, for richer kinds of computation, the empirical evidence suggests that there are many shades of expressivity, which is why no-one has ever postulated an analogue of the CTT for them. Therefore this paper makes a radical contribution which attempts to overturn 80 years of misguided extrapolation, namely, that accepting the CTT implies that all programs that are of interest to computer science, cognitive science, and philosophy have nothing else but the expressive capabilities of input-output processes.

In conclusion, we argue that an incorrect interpretation of the CTT underlies the persuasiveness of the CRA because it allows Searle to argue that the limitations of the program he presents in the CRA applies to all programs. Accepting this generalisation strategy, as Searle does, means there can be no syntactic formal processes in a program, of even fiendish complexity or strangeness, that will ever give rise to any kind of semantics. This paper agrees with Searle only insofar as when programs are confined to using input-output computation and so are within the expressivity-scope of a TM, that there can be no meaning in the internal symbol processing. For programs outside the expressivity-scope of a TM, meaning may appear due to processing occurring that is more computationally expressive than that of a TM. For the CRA to work it must apply to all programs. But we have shown it cannot use the CTT to generalise like this.

Whilst more expressive or powerful formalisms may natively sup-

³ The aim of hypercomputation is to find input-output processes that compute functions that cannot be realised with a Turing machine.

port more complex computational processes that are basic to that formalism, this does not mean that such formalisms can carry out function computations that input-output computation cannot carry out. It is, of course, possible to use a rich interactive machine to implement a simple function; after all, that is what we do with our modern computers all the time. It is our belief that this does not lead to new computable functions, i.e., some sort of “hypercomputation”. In other words, the limits with regards to computation established by the CTT are valid even if more sophisticated machinery than a TM is employed. It is the other direction that is the core of this paper: When considering more sophisticated computational tasks, then standard Turing machines (and their mode of operation) are not sufficient to explore the range of possibilities.

REFERENCES

- [1] J. Cabessa and H.T. Siegelman, ‘The super-turing computational power of plastic recurrent neural networks’, *International Journal of neural systems*, **24**, 450029–450051, (2014).
- [2] Matthias Felleisen, ‘On the expressive power of programming languages’, *Science of Computer Programming*, **17**, 35–75, (1991).
- [3] Cosimo Laneve and Antonio Vitale, ‘The expressive power of synchronizations’, in *2010 25th Annual IEEE Symposium on Logic in Computer Science*, pp. 382–391, (2010).
- [4] J.R. Searle, ‘Minds, brains, and programs’, *Behavioral and Brain Sciences*, **3**, 417–424, (1980).
- [5] J.R. Searle, *The Rediscovery of the Mind*, MIT Press, Cambridge, 1992.
- [6] H.T. Siegelman, ‘Computation beyond the turing limit’, *Science*, **268**, 545–8, (1995).
- [7] A. Sloman, ‘Beyond turing equivalence’, in *Machines and Thought: The Legacy of Alan Turing (vol I)*, eds., P.J.R. Millican and A. Clark, 179–219, The Clarendon Press, Oxford, (1996).

The relationship between goals and beliefs, and some implications for computation

Nathaniel Virgo¹

Abstract. We discuss a recent mathematical framework for understanding the relationship between a system being *able to perform a task* and that system *having a model* of its environment. We argue that this framework has much to say about the notion of computation, both in its own right and in relation to cognitive science. In particular, it suggests a reason for the close association between computation and *semantics*, in the sense of a mapping from symbols to their meaning. Along the way, we raise an interesting, though speculative, hypothesis about what’s happening inside the brain when we experience a sudden moment of insight.

1 Introduction

Recently, my colleagues and I have been working on a mathematical framework for understanding the relationship between a system being *able to perform a task* and that system *having a model* of its environment. Such a connection was proposed by Conant and Ashby [5] in the early days of cybernetics, and our work is part of a tradition of building on the idea (e.g. [8, 10, 9]; see [13] for more on how our own work fits in). In common with most of these results, our work is mathematically only distantly related to Conant and Ashby’s original. Our full framework is not yet publicly available, but we describe a cut-down version in [13] that includes many of the main ideas.

In this discussion article, I will describe the framework informally, and talk about some of its implications for the notion of *computation*, in both biological and artificial systems. (This was the topic of the Foundations of Computation Symposium at AISB 2025, at which I presented this work.) The framework is discussed in section 2, and its implications for computation and cognitive science are discussed in section 3, along with a tentative hypothesis about what happens when we experience a sudden moment of insight.

2 A description of our framework

Our framework is designed to capture the idea of a system that behaves intentionally, interacting with its environment in such a way as to achieve some specific aim. To think about this systematically, we first consider coupled dynamical systems, a familiar idea in cognitive science. In the most basic setting we would have a pair of systems, the agent and its environment. [13] treats this case. However, as we will see later on, it is often worthwhile to be more fine-grained than this, dividing either the agent or its environment into multiple sub-systems. This also allows us to consider the question of *where to draw the boundary*, since we can vary which sub-systems are considered part of the agent and which ones part of its environment. In the

full framework, we achieve this using the tools of categorical systems theory, making particular use of the notion of a *behavioural theory* [12, chapter 3].

As with the rest of the framework, I will treat dynamical systems and their coupling informally in this discussion. It is worth saying, however, that our framework, at least in its current form, works with *possibilistic* dynamical systems, in which there may be many possible next states. We say “possibilistic” rather than “probabilistic” because we are only concerned with which states are possible given the current state, and not how likely they are to occur.

The notion of coupling we use is one of mutual constraint: each system has an interface, and when two systems are coupled they can only undergo transitions that agree on the interface. An image to have in mind is that of two people with their hands stuck together: they have substantial freedom to move their hands around, but they can only do so in such a way that each person’s hand is always in the same position as the other’s. The coupled system thus has less freedom than its two subsystems put together (i.e. the same two people, but with their hands separated). The whole is less than the product of its parts, in that the set of trajectories available to the coupled system is a subset of the Cartesian product of the sets of trajectories that would be available to its two components if they were separate.

A central feature of our approach is that, we claim, in order to talk about intentional dynamics, it isn’t enough to specify the dynamics in question: one also needs to specify what the dynamics are supposed to be trying to achieve.

This is quite a subtle point. It might seem, *a priori*, rather questionable to consider purpose or intention in the context of biology, and so we will cover that point below. In the case of an artificial system, however, it will often seem a very reasonable thing to consider. When building a control system, we usually do have some kind of goal in mind: if we are building a component of a flight control system, for example, we might specify that the controller must be capable of maintaining level flight, along with a rather strict definition of what counts as level flight and in what conditions it must be maintained. This specification is not part of the autopilot itself, and cannot necessarily be determined from examining the flight controller alone. Nevertheless, examining the controller in the context of knowing what it was designed to do might yield insights that would be hard or impossible to obtain were one to regard the controller strictly as a dynamical system, without any intentional interpretation attached. (If this is not clear now, we hope it will become so later.)

Formally, in our framework, such goals are specified in terms of subsets of a system’s state space. This might be the set of all states of the aircraft in which its attitude is within some specified margin of its target attitude, for example. In this example, the goal is that when the aircraft is coupled to the controller in the appropriate way, and when

¹ University of Hertfordshire, England
n.virgo@herts.ac.uk

the two systems are appropriately initialised, the aircraft’s attitude will remain within this margin indefinitely. In another example, the goal might be to *reach* a given end state, starting from an initial state that is outside it.

Reasoning in terms of subsets of the state space in this way is a standard approach within control theory. The main point here is that the specification of the goal is *additional data*—it’s not part of the dynamics and needs to be given alongside them in order for our framework to be applied.

Let us now return to the questions of goals and purpose within biology. It would certainly be naïve to imagine that biological features have a specification of their design goals in the same way that an autopilot would; evolution simply doesn’t work that way. Nevertheless, in some situations and for some purposes it can be helpful to think of biological features as being ‘for’ something. Understanding the heart *as a pump* tells us a lot about what it does and how it behaves, even if it doesn’t tell us everything. This is essentially Dennett’s idea of the *design stance* (e.g. [6]), at least on one possible reading, and in general we take a Dennettian approach to ideas about goals or purpose in biology. To apply our framework to a biological system, one must supply both the dynamics of the system in question (or at least a model of them), and a specification of its assumed goal. But this goal need not be seen as a fact or even a hypothesis about how the system came into being; it is merely a specification of a goal the user *wishes to regard the system as having*. This is a choice that a user of our framework must make, and for a given system it is likely that some choices of assumed goal will lead to more insightful conclusions than others.

We regard this goal-dependence as a feature rather than a bug: it forces us to say explicitly that we’re attributing a goal to the system, as well as exactly what goal it is, rather than allowing it to be tacitly implied. Our philosophical position (or rather mine, at least) is that whenever we speak of one system having knowledge (or beliefs, or a model, or a representation) of another, it is always in the context of an assumed goal. We aim to make the relationship as explicit as possible.

It is worth mentioning one specific approach to goals in biology, namely the idea that an organism’s goal is ultimately to maintain its own existence. An early version of this idea appears in Ashby’s [1], which discusses the idea that an organism has a set of essential physiological variables and a *boundary of viability*—a region of values for these variables for which the organism is alive and not in immediate danger—and that an organism’s behaviour is largely driven by the need to remain inside this viable region. This seems to have influenced Maturana and Varela in [11], who saw an organism’s need to maintain its own viability as a defining characteristic of life. The notion of viability boundary became a central concept in enactive cognitive science, particularly in works focused on modelling such as [2, 3, 4, 7]. This notion is easily accommodated within our framework, since the viable region is simply a subset of the states of a system.

It is worth noting, though, that this is in some sense a dual of the control theory setup. In a typical control problem, the goal relates to a subset of the states of the *environment*, or plant, while in this enactive perspective on biology, the goal relates to a subset of states of the *agent*: the ‘controller’ is controlling itself, rather than some external system. Nevertheless, the agent may well have to engage with its environment in order to succeed at the task of maintaining itself. Even if our only goal is to maintain our blood sugar level, we still have to interact with our surroundings in all manner of non-trivial ways in order to be able to eat. Our framework treats these

cases of *intrinsic* and *extrinsic* regulation in very similar ways.

The interesting feature of our framework is a result that generalises the “good regulator theorem for embodied agents” that we described in [13]. Roughly, the result says that whenever we have an agent that is capable of performing some control or regulation task (whether intrinsic or extrinsic), we can always *interpret* its internal state as encoding *knowledge* about the state of its environment. A state of knowledge (also known as a belief state) is a subset of the environment’s states, to be thought of as the set of states that the agent thinks the environment might be in. To interpret the agent’s state as encoding knowledge simply means that we have a function $X \rightarrow \mathcal{P}(Y)$ that obeys some axioms, where X is the set of states of the agent and $\mathcal{P}(Y)$ is the set of subsets of the environment’s state space. The axioms ensure that the belief state updates in a logical way when the agent gets new information: if it is possible for the environment to have transitioned to some new state y given the information that the agent received, then y should be included in the new belief.

We then find that the agent’s actions are consistent with these beliefs, in that any action the agent actually takes is consistent with an action an idealised ‘rational’ agent might take, if it indeed had this belief about the environment’s state. Thus, although we started with an agent that was specified dynamical system, with no particular meaning attached to its states, it can nevertheless be interpreted ‘as if’ it were behaving rationally, and its states take on a semantic interpretation that arises naturally from the framework.

The way in which this belief interpretation is constructed is not complicated. It is described in [13], and we also outline it in regards to the example in section 3.1.

3 Implications for computation

The main point of this discussion article is to relate our framework to the notion of computation. There are many possible ways to define computation, corresponding to different perspectives on what the term should mean. Here, at the level of a slogan, we will say that what makes something a computation rather than mere dynamics is that *a computation has a right answer*.

As with the rest of our framework, this might initially seem counterintuitive, because it brings in a notion of purpose, which one might feel doesn’t belong in a mathematical description. But we find that, as for control systems, a suitable notion of end goal is not hard to formalise, and doing so seems to give more insight than one might expect.

In particular, the study of computation is intimately tied to the notion of semantics. At the physical level the operation of a computer can be described in terms of electrical currents and voltages, but we almost always reason about computations in terms of symbols that have a specific semantic interpretation. Anyone who has programmed a computer has stored intermediate data for a computation in a variable and reasoned about what the contents of that variable *mean*, in terms of the computation that’s being performed. This is quite different from how we reason about physical or dynamical systems that aren’t computational in nature, where we reason about how things change over time without ascribing any particular semantics or meaning to them.

I claim that these two aspects of computation—that a computation has a right answer, and that the notion of meaning seems to play a pervasive role—are closely related. In the following simple example, the notion of meaning of a variable arises automatically by applying our framework to a computation. We suspect that the strong connection between computation and meaning or semantics arises in general

for the same reason.

Although simple, the example itself suggests some interesting, though speculative, hypotheses about cognitive science. After the example, we will comment on broader implications of our approach for cognitive science.

3.1 An example, and a hypothesis about thinking

Let us examine a very simple example of a computation, using our framework, but in a very informal way.

To do this, we must specify not only the system doing the computation (the ‘agent’), but also its environment, and what it means for the computation to return a correct answer. The example we will choose is an agent that must add two numbers.

We start with the environment, since this is rather simple: first, it picks two natural numbers and sends them to the agent. These numbers are picked possibilistically, meaning that they can be anything, with no probability distribution associated to them. (An actual environment would use some mechanism to pick them, which might or might not involve some randomness and hence a probability distribution. But we desire our computation to return the correct result for *any* two numbers, and because of this we model the environment as being able to pick any two numbers arbitrarily.)

The environment then calculates the sum of the two numbers, which we assume happens instantaneously, again because we are not really concerned with the details of the environment. It then waits for a response from the agent, which can take an arbitrary number of time steps. Finally, if and when the agent returns a result, the environment checks whether the agent’s response is equal to the sum of the two numbers that it stored earlier.

The *goal* is for this check to return true, i.e. the agent is supposed to calculate the sum of the two numbers. This goal is a formal piece of data that has to be specified in order to perform our analysis.

We will think of the agent as a simple electronic device, and assume it works as follows. Inside the agent are two registers, register *A* and register *B*, which can each hold a single natural number.² To ease the exposition, we will assume that some of the agent’s circuitry is associated with register *A* and some with register *B*; this is a choice that we have made about where to draw the boundary between two sub-systems of the agent.

When the agent receives the two numbers from the environment, they are placed into the two registers. Then, at each time step, the following sequence of events occurs:

- first, register *A* checks whether its value is 0. If it is, then register *A* sends a special `DONE` message to register *B*, and register *B* returns its contents to the environment.
- otherwise, register *A* decrements its value by 1, and register *B*, having not received the `DONE` signal, increments its value by 1.

Thus, register *A* will count down while register *B* counts up, until register *A* gets to 0, at which point register *B* contains the sum of the inputs, which will be sent to the environment.

We can now say what each register knows about the state of the environment and the other register. We emphasise that although we are presenting it somewhat informally, this talk of ‘knowledge’ is not informal at all: it is something that we calculate formally from the specification of the dynamics and the goal.

² This means each register can hold a potentially infinite amount of information. If the reader finds this concerning, they are free to add appropriate upper bounds throughout the example. This can be done without substantially changing the analysis.

The calculation for register *A* works as follows: given the value of register *A* (a natural number) we find the set of all pairs

(state of environment, state of register *B*)

(i.e. pairs of natural numbers) such that the agent-environment system will eventually enter a state where the environment checks the agent’s answer and returns true. The calculation for register *B* is similar, with *A* and *B* reversed.

Writing *a*, *b* and *c* for the numbers stored in register *A*, register *B* and the environment, the result of this calculation are as follows: first, when register *A* is in state *a*, its knowledge state consists of all those pairs (*b*, *c*) such that $a + b = c$. This means that register *A* never knows the value of *c*, but it does know that $a + b = c$. In particular, when the value of *a* is 0, register *A* doesn’t know the value of *c*, but it does know that register *B* currently contains the correct answer, i.e. $b = c$.

For most of the execution time, the situation for register *B* is similar: it doesn’t know the value of *c*, but it does know that $a + b = c$, i.e. its belief state is the set of pairs (*a*, *c*) such that $a + b = c$. The exception to this is when register *B* receives the `DONE` signal. After receiving the signal, register *B* (together with its associated circuitry, with its own additional state for keeping track of whether the signal has been received) knows that $a = 0$, and hence it knows the value of *c*: it’s equal to its own state, *b*. It is at this point that register *B* returns the result.

There are several things to note about this example. We can start by noting that the analysis is quite intuitively reasonable: our claims about the two registers’ knowledge seem to be quite close to what one might think about if one were designing the circuitry associated with the two registers. This is an outcome of our analysis rather than an assumption.

However, this is in part due to our having made appropriate choices about where the boundary should be drawn; we can imagine that if we had drawn it in a different place, for example by associating all the circuitry with one of the two registers, then the analysis would have seemed much less intuitively reasonable. The question of how to draw the boundaries in the right place is a subtle one, and we suspect that fully resolving it mathematically will be equally subtle.

In particular, we could have chosen not to draw an internal boundary at all, instead considering both registers and all the circuitry together as a single system. Such an analysis provides much less insight, at least when considered in isolation. Since we always have $a + b = c$, the value of *c* can always be deduced from the value of the pair (*a*, *b*). Because of this the analysis says that the agent ‘knows’ the sum of its two inputs as soon as it receives them. It then waits for a time before returning the answer. This delay is no problem, since the task doesn’t put any constraint on when the answer should be returned, but without dividing the agent into subsystems, the analysis says nothing about what’s happening before the result is returned.

We can draw a rather speculative analogy between this phenomenon and a human’s need to think before answering a question. We may sometimes say a person knows something, even if they might have to think a little while before being able to properly recall it. It seems reasonable to propose that this situation is similar to our example: the knowledge is “already there in the person’s head”, but needs to be processed first, before it can be said. Perhaps the reason for this need for processing is also analogous to the example: in order for the thing to be said, the information needs to be present not just in the brain as a whole, but in the specific parts of the nervous system that are responsible for forming sentences and translating them into muscle movements in the vocal tract. Moreover, the information

needs to be present in the right form such that these tasks can be performed. It seems very reasonable that the need for thinking before recalling something might be driven by this, at least in part.

This suggests to us an even more speculative hypothesis. In the example, prior to receiving the DONE signal, register B knows that $a + b = c$, but it doesn't know the values of a and c themselves. Thus, its knowledge about c in particular consists only of a lower bound, $c \geq b$ (since a is known to be nonnegative). However, after receiving the DONE signal, it knows this lower bound is attained; its knowledge collapses in a single time step from an infinite set of possibilities to a single one. This puts us in mind of the moments when we experience a sudden flash of insight, when the answer to a problem “suddenly comes to us” after thinking about it for a long time. We can speculate that what we're experiencing here is similar to the example: a computation (for want of a better word) has been in progress, and as a result of the computation the answer is already present in that part of the brain that is able to form sentences and report about knowing the answer—but this part of the brain does not yet know *that* it is the answer.³ The very tentative hypothesis is that what happens during the ‘sudden moment of insight’ is that an analogue of a DONE signal is received. The relevant part of the brain goes from having all the information about the answer present in the correct form, to additionally knowing that this information is indeed the answer to the problem, and hence being able to report on it.

We do not know how such a hypothesis could be tested, but it seems interesting enough to report on, in the hope that others in the neuroscience and consciousness communities might have some insight.

3.2 Some tentative implications for cognitive science

Our example involved an extrinsic goal: the goal of the program was to return the correct answer to the environment. This seems to us somewhat inherent in the notion of computation, at least as we understand that word within computer science. However, we can imagine similar situations arising within the boundary-of-viability paradigm, where an organism must shuffle information around inside itself in a similar way, in order to achieve its own intrinsic goal of self-maintenance. If such a process takes time then we might refer to it as ‘thinking’, and indeed I believe that on some level this is exactly what thinking is.

Does this mean that thinking is computation? I think this is to some extent a matter of definition. The distinction between intrinsic and extrinsic goals is real, and might be enough to warrant not mixing the two terms.

But perhaps a more important difference is that, usually, in the case of a computation the goal is something very specific that was pre-specified before the program was written. In contrast, if we take the Dennettian approach to goals in biology that we outlined above, biological goals are much less ‘real’, and much less precisely specified. They represent an attitude of the person studying the organism, much more than a real feature of the organism itself.

Thus, when we refer to a biological process in computational terms (including when we say its states ‘represent’ knowledge or beliefs), I think we are *imposing a goal on the system*, and reasoning

about its dynamics in terms of that assumed goal. I don't think this is necessarily a bad thing to do—in fact I think it can give a lot of insight into cognition and other biological processes—but I also think it's important to realise that this is what we're doing, and to make our assumed goal as explicit as possible. A framework that makes these ideas mathematical could therefore be quite valuable.

ACKNOWLEDGEMENTS

This work is funded by the Advanced Research + Invention Agency (ARIA) through project code MSAI-SE01-P011.

REFERENCES

- [1] W. Ross Ashby, *Design for a brain*, Wiley New York, 1960.
- [2] Randall D. Beer, ‘The dynamics of adaptive behavior: A research program’, *Robotics and Autonomous Systems*, **20**(2-4), 257–289, (1997).
- [3] Randall D. Beer, ‘Autopoiesis and Cognition in the Game of Life’, *Artificial Life*, **10**(3), 309–326, (2004).
- [4] Randall D. Beer, ‘The cognitive domain of a glider in the game of life’, *Artificial life*, **20**(2), 183–206, (2014).
- [5] Roger C. Conant and W. Ross Ashby, ‘Every good regulator of a system must be a model of that system’, *International journal of systems science*, **1**(2), 89–97, (1970).
- [6] Daniel C. Dennett, ‘Intentional systems’, *Journal of Philosophy*, **68**(4), 87–106, (1971).
- [7] Ezequiel A. Di Paolo, ‘Autopoiesis, adaptivity, teleology, agency’, *Phenomenology and the Cognitive Sciences*, **4**(4), 429–452, (December 2005).
- [8] Bruce A. Francis and Walter Murray Wonham, ‘The internal model principle of control theory’, *Automatica*, **12**(5), 457–465, (1976).
- [9] Karl J Friston, ‘A free energy principle for a particular physics’, *arXiv preprint arXiv:1906.10184*, (2019).
- [10] Artemy Kolchinsky and David H. Wolpert, ‘Semantic information, autonomous agency and non-equilibrium statistical physics’, *Interface Focus*, **8**(6), 20180041, (2018).
- [11] Humberto R. Maturana and Francisco J. Varela, *Autopoiesis and cognition: The realization of the living*, Springer Science & Business Media, 1980.
- [12] David Jaz Myers. Categorical systems theory. Online book draft <http://davidjaz.com/Papers/DynamicalBook.pdf>. Accessed: 2025-05-07, 2023.
- [13] Nathaniel Virgo, Martin Biehl, Manuel Baltieri, and Matteo Capucci, ‘A “good regulator theorem” for embodied agents’, in *ALIFE 2025: Ciphers of Life: Proceedings of the Artificial Life Conference 2025*, pp. 419–428, (10 2025).

³ By “part of the brain”, we don't mean to imply some kind of *module* whose *purpose* is to do this in any pre-given or absolute sense; we are merely hypothesising that some regions could productively be considered subsystems in the sense of our framework, such that the story we are telling would make sense within our framework.

Artificial Autopoiesis

Steve Battle¹

Abstract.

What is gained, and is anything lost, by treating cognition, life, and other processes computationally? Viewed within the strict confines of the Church-Turing thesis, computational approaches address only effective procedures. Computational modeling enables simulations – such as those described here – that provide insight into living (autopoietic) processes, but misses one key aspect, namely interaction with the physical environment. Cybernetics provides us with a set of tools that is complementary to computational thinking, where the focus is precisely this interaction with the environment. The approach presented here combines computational modeling with cybernetic analysis of the results.

1 Introduction

Autopoiesis, meaning “*self-creation*,” is a term coined by Chilean biologists Humberto Maturana and Francisco Varela to describe systems capable of maintaining themselves by creating their own parts from materials existing in the physical milieu [6]. Understood as a definition of the primary characteristics of living systems, features such as DNA, reproduction and evolution are regarded as contingent to biological life. By removing our biological blinkers, it raises the fascinating prospect of being able to identify extra-terrestrial or artificial life-forms as such.

Life, as defined by Autopoiesis, must satisfy just two conditions:

- **Organisational closure** of a set of processes of production. An autocatalytic set of entities that can catalyse their own production.
- **Structural closure.** Something like a cell wall to partition the system from its environment.

Autopoiesis operates only in the physical domain. We can build computational models of living systems, but these can never be more than simulations. Overcoming this hurdle awaits the development of self-organising nano-machinery. In the meantime, simulation can teach us a lot about autopoiesis; not real —but *artificial autopoiesis*.

Francisco Varela further clarified the nature of autopoietic systems [9] by presenting a simulation based on a cellular automaton [4]. This defines a small set of reaction rules similar to chemical reaction rules. In Figure 1, a catalytic agent (asterisk) transforms substrate particles (circles) in the environment into its structural building blocks (squared circles). These are able to self-organise, linking themselves into a structure akin to a long-chain polymer. This chain of bonded particles can close in on itself to form a closed boundary, analogous to a cell wall, enclosing the catalyst.

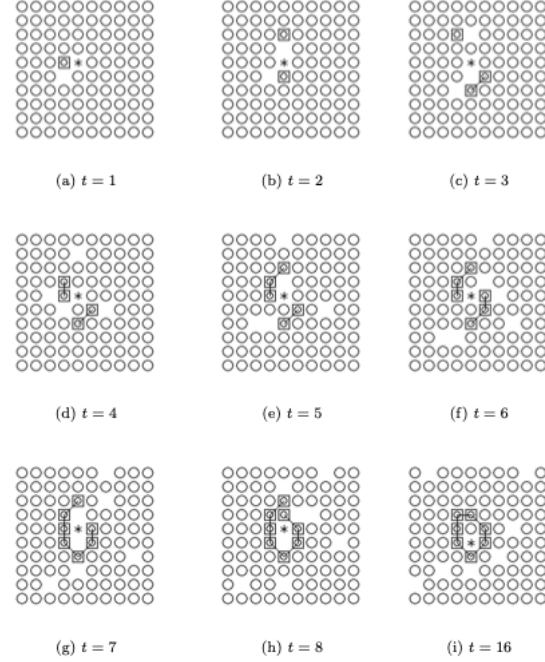


Figure 1: Cellular automaton simulation of autopoiesis based on Varela’s original algorithm. In (a) a pair of substrate (circles) are transformed into a single link particle (squared circle) by the catalyst (asterisk). By (c) we see the first bonds forming, then in (i) these form a closed boundary around the catalyst.

Varela’s reaction rules introduce a substrate, S , of particles that interact with a catalyst, K . When a pair of substrate particles collide with the catalyst they are composed to form a link particle, L , leaving the catalyst unchanged. Link particles are so named because on mutual contact a pair of link particles become concatenated. Link particles decay over time, randomly disintegrating back into a pair of substrate particles from whence they came. A reaction rule is shown as a transformational arrow that defines the reactants on the left-hand side, with the reaction products to the right of the arrow. A plus symbol between reactants indicates a collision event where the particles come into contact. The particle type may be prefixed by a number indicating an integer number of particles of the same type. The equivalent of chemical bonds are indicated by multiplicative operators such as a product or exponentiation.

composition: $K + 2S \longrightarrow K + L$

concatenation: $L^n + L \longrightarrow L^{n+1}$

disintegration: $L \longrightarrow 2S$

¹ School of Computing and Creative Technologies
University of the West of England, Bristol, UK.
Steve.Battle@uwe.ac.uk

To enable smoother simulation, we switch from a cellular automaton to a *liquid automaton* [2]. Liquid automata do away with discrete space-time, being based instead on 2D physics engines such as Box2D [3]. This allows particles to move around freely and sometimes collide with each other, at which point a reaction may take place if the right conditions are met.

Using Varela’s original set of reaction rules in a liquid automaton produces behaviour that is qualitatively equivalent to Varela’s cellular automaton. While sporadically achieving structural closure, these structures prove to be extremely fragile. In Figure 2 we see an enclosure that has formed around the catalyst, but there are also random cast-offs; chains that failed to close that haven’t yet disintegrated. Even the enclosure that has formed will break as soon as any one link in the chain disintegrates. As substrate floods in through the gap, newly formed links may repair the hole, but equally it may be flayed away from the catalyst by the bustling activity of the substrate.

To enable further analysis of the simulation, we need to increase the stability of the cellular enclosure. Cell membranes are like soap bubbles that form spheres to minimize their surface area. We require a similar approach that will naturally form a circle to minimise its circumference in 2 dimensions. The simplest amendment we can make to Varela’s rules is to link the composed particles directly to the catalyst at a fixed radius rather than to each other.

We use an amended set of reaction rules that produce more stable results when implemented in a liquid automaton. The rules introduce a substrate, S , of particles that interact with a catalyst, K . When a pair of substrate particles collide with the catalyst they are composed to form a link particle, L , that remains linked to the catalyst, denoted by the product KL . As before, Link particles disintegrate over time.

composition: $K + 2S \rightarrow KL$

disintegration: $L \rightarrow 2S$

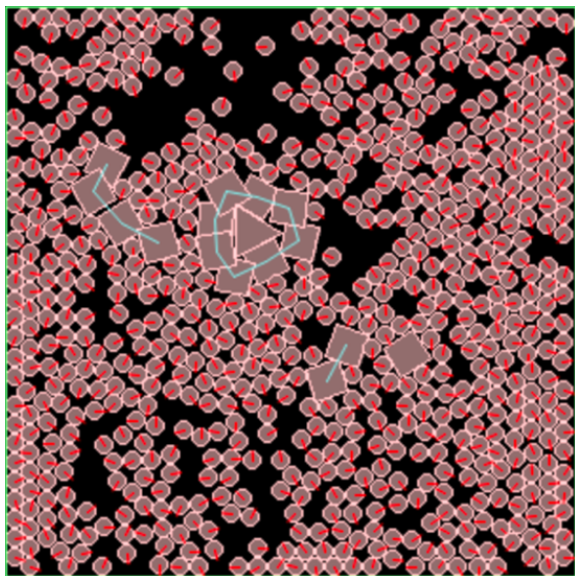


Figure 2: Varela’s reaction rules running in a liquid automaton. This screenshot shows a point where Link particles (squares) have formed an enclosure around the catalyst (triangle).

The effect of bonding Link particles to the catalyst at a fixed radius is that they naturally form a circular enclosure around the catalyst as

simulated in Figure 3. As link particles are produced, they punch their way into the existing wall and spread out to form as near a circular enclosure as possible. Furthermore, the enclosure is much more stable as it doesn’t lose integrity with the loss of a single link.

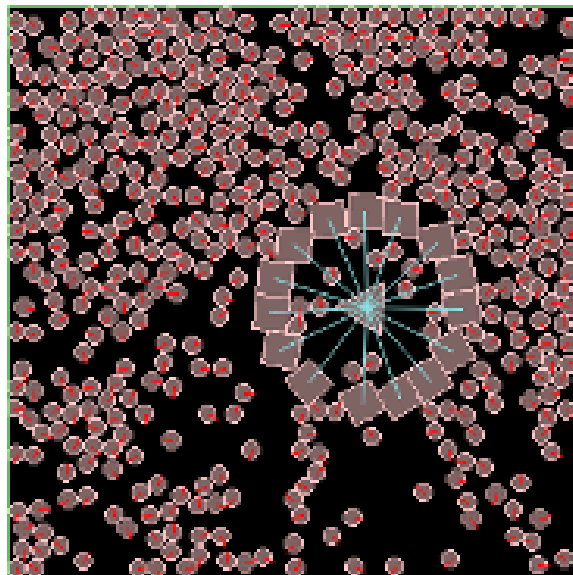


Figure 3: Modified reaction rules produce a stable cell-wall (structural closure). Substrate particles, S , are circular. Spokes (blue) connect the square Link particles, L , to the triangular catalyst, K .

2 A cybernetic approach

In seeking explanations of the behaviour of living systems, the simulation of the previous section offers a guide. The underlying computational system is an organisationally closed set of reaction rules. However, to understand this as a simulation of a living system is to ask, “*what does it do?*” In other words, what is the overall effect of these reaction rules working together as a whole? Cybernetics focuses not on the underlying computation but directs us to look for regulatory, feedback processes [10]. Can we measure some property of the system and show that this is being actively controlled; this is the controlled variable [5]. From a cybernetic perspective we would also say that this is the purpose of the system [7]. We must also investigate how the system gathers information about the controlled variable, and by what means the system is able to influence the controlled variable [1]. There are many system variables we can measure:- the number of substrate or link particles, the number or density of particles within or without the radius of the cell wall.

From an autopoietic perspective the function of the organisation (the reaction rules) is to maintain structural closure. The resulting enclosure defines a distinction between self and the environment, that in turn implies a relationship between the inner and outer realms. The wall is not simply a shield to keep the world out, but is semi-permeable, allowing sufficient resources into the cell to maintain its existence. In this simulation, permeability is achieved simply by gaps in the wall.

There is a circular relationship between the link particles in the cell wall and the flow of substrate particles to the catalyst. The permeability of the enclosure to substrate particles is directly related to the number of link elements and the radius of the cell wall. We observe that the density of substrate within the cell-wall is lower than

outside. This lower density leads to a lower rate of composition and so the production of link particles is reduced. With the disintegration of link particles, the cell-wall becomes more permeable leading to a corresponding increase in the production of links.

The composition rule produces new link particles, L , and this is understood as how the system influences its environment. One of the reactants involved are the substrate particles, S , so this enables the system to gather information about the quantity of reactants in the vicinity of the catalyst. We therefore expect the density of link particles within the cell wall to achieve a dynamic equilibrium.

We understand this as a negative feedback loop in which the error is conveyed via the flow of substrate to the catalyst. However, the quantity of substrate colliding with the catalyst can only be a positive integer so this is an example of a positive-error-only feedback loop. This kind of feedback acts unidirectionally, where the control system can only correct for under-shoot. In the case of over-shoot, the required reduction in the number of link particles must await their ‘natural’ disintegration, but this occurs at a constant rate and is blind to the total number of link particles so is independent of feedback.

The relationship between simple organisms and their environment concerned Herb Simon [8] who recognised that an organism “needs only to maintain a certain average rate of food intake, and additional food is of no use to it. In the psychologist’s language, it has a definite, fixed aspiration level, and its successes or failures do not change its aspirations.” In other words, the artificial autopoietic simulation in Figure 3 ‘aspires’ to construct an enclosure by feeding on the substrate.

3 Results

Based on 150 runs, each 200 seconds long, we collect data on the population size, the cell radius, the number of link particles, and the number and density of substrate particles within and without the cell. On a typical run, as seen in Figure 4, the number of Link particles, L , produced by the composition rule achieves a dynamic equilibrium as does the quantity of substrate, S_{in} , within the cell wall. We expect the cell-wall to regulate the density of substrate within the cell wall, that is, to maintain this at a lower level than in the external environment. A 1-sided t-test to test this directional hypothesis confirms that that this difference exists ($t = 6.0888E - 191, p < 0.001$).

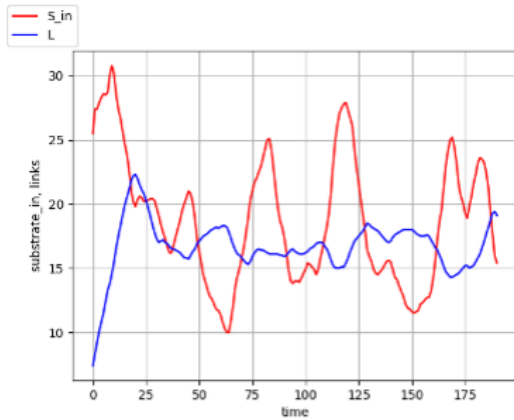


Figure 4: Given an initial population of 500 substrate particles and a cell-radius of 6, the number of link particles, L (blue), achieves a dynamic equilibrium of 17.07. The quantity of substrate within the cell-wall, S_{in} (red), with a mean of 17.88, rises when the number of links drops and it becomes more permeable.

On every run the initial population of substrate is varied randomly between 450 and 500, and the cell radius from 4 to 9 units (where the diameter of a substrate particle is 1 unit). We first determine which of the disturbances the system can accommodate, and which result in shifting the goal.

The results can be seen in the sensitivity analysis, Figure 5, showing what fraction of the output’s variance is explained by each input alone. If there were no functioning cell wall, as would result from broken organisational closure, then the production of links, L , would be limited only by the total quantity of substrate available. However, this is not the case, and we see weak sensitivity to this (0.026). For the cell radius we see much greater sensitivity (0.894). For this reason, we must consider the correlation between the controlled variable (substrate density inside the cell) and the radius.

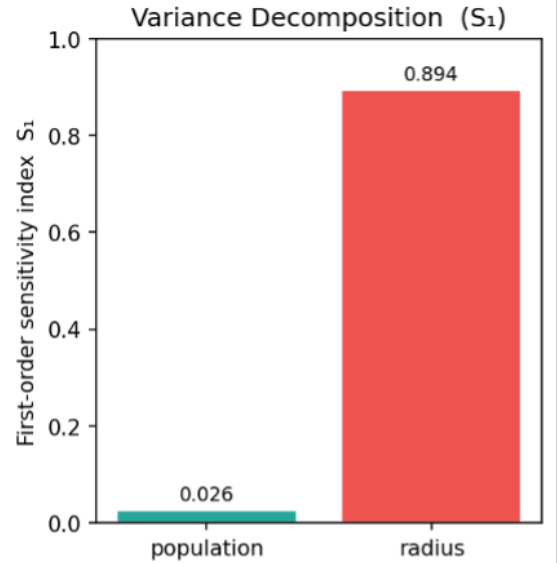


Figure 5: Sensitivity analysis determines which of the disturbances have an effect the controlled variable. Varying the initial population of substrate has little effect, but changing the cell radius has large effect.

The goal of the system is largely independent of the initial population of substrate but is highly dependent on the length of the catalyst/link-cell radius that determines the circumference of the cell-wall that it needs to construct.

Consider the linear relationship between the output of the system (link particles) and the radius of the cell. The Pearson correlation between the number of link particles and the cell radius, $r = +0.989(p = 1.036e - 124)$. Therefore, the number of particles produced is effectively a function of the cell-radius as seen in Figure 6.

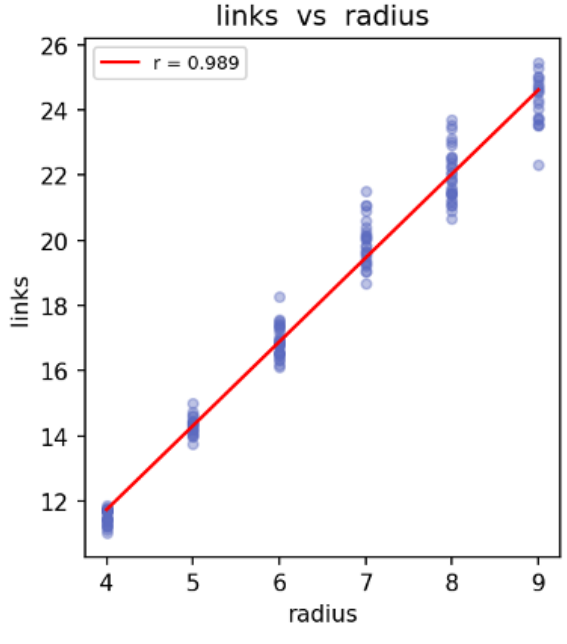


Figure 6: The mean number of links correlates highly (0.989) with the cell radius; they are in a dynamic equilibrium that provides isolation for the cell, while allowing sufficient permeability.

Next, consider the mean density of substrate inside the cell. The Pearson correlation between the density of substrate within the cell and the radius, $r = -0.841$ ($p = 3.676e - 41$).

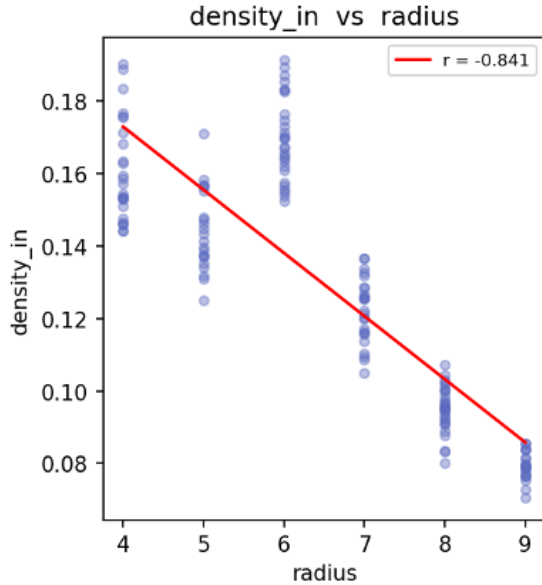


Figure 7: The density of substrate within the cell wall is (negatively) correlated (-0.841) with the radius. The goal varies slightly with the radius but is relatively stable, with the number of links in the cell-wall providing an effective means of control.

The linear plot ($density_{in}$ vs $radius$) in Figure 7 shows the effective target goal for the substrate density inside the cell as the cell radius is varied.

4 Conclusion

Computational modelling is an aid to building artificial autopoiesis, used here to simulate an organizationally closed set of reaction rules. The structural closure that emerges is best understood using the methods of cybernetics. This approach enables us to identify the purposeful behaviour of the system, where the focus is the relationship between the self and its environment.

REFERENCES

- [1] W. Ross Ashby, *An Introduction to Cybernetics*, Chapman & Hall, 1956.
- [2] Steve Battle, 'Simulated autopoiesis in liquid automata', *Int. Journ. of Unconventional Computing*, **19**(2-3), 137–147, (2026).
- [3] Erin Catto. Box2D: A 2d physics engine for games. <https://box2d.org>, 2026. Accessed: 2026-07-15.
- [4] Martin Gardner, 'MATHEMATICAL GAMES', *Scientific American*, **223**(4), 120–123, (1970).
- [5] R S Marken, *The Study of Living Control Systems: A Guide to Doing Research on Purpose*, Cambridge University Press, 2021.
- [6] H R Maturana and F J Varela, *Autopoiesis and Cognition: The Realization of the Living*, D. Reidel Publishing Company, 1980.
- [7] Arturo Rosenblueth, Norbert Wiener, and Julian Bigelow, 'Behavior, Purpose and Teleology', *Philosophy of Science*, **10**(1), 18–24, (1943).
- [8] H.A. Simon, *Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in Society Setting*, Wiley, 1957.
- [9] Francisco Varela, *Principles of Biological Autonomy*, Elsevier North Holland, Inc., 1979.
- [10] Norbert Wiener, *Cybernetics: or Control and Communication in the Animal and the Machine*, MIT Press, Cambridge, MA, 1948.

Comparing the organization of information in mechanical and biological systems

Avel Guénin–Carlut¹

Introduction

The cognitive sciences inherited a foundational analogy from the Macy conferences: the mind as software implemented on the computational architecture of the brain [11]. The strong version of the analogy, exemplified by Chomsky [9] or Fodor [19], fell out of mode in the 80s and 90s due to the generally recognized failure of symbolic robotics [7, 8] and the progress of “connectionist” architectures reconstructing cognitive function from neuronal modelling [35, 12]. Nonetheless, its influence can still be felt in the largely successful program of *cognitive neuroscience*, which seeks to explain cognition as the product of specialized information processing mechanisms in the brain [1, 28, 29]. In this approach, *computation* emerges from the systematic way in which the brain *represents* incoming signal and *processes* these representation. However, the underlying architecture is not assumed to constitute a simple system of logical instructions which are somehow implemented into the architecture of the brain. Rather, it is understood to be a complex process, embedded in the specific biophysical architecture of neurodynamic organization at all relevant scales of activity.

Nonetheless, some would decry the mechanistic, computational approach to cognition as being either reductive or outright incoherent with the actual dynamics underlying cognition. Roli et al. [30] suggest that the specific self-constructive aspect of biological systems drives an open-ended evolution of structure and meaning, which cannot be reproduced by a *bona fide* computational system. Jaeger et al. [26] argue that the realization of relevance is by construction grounded in biological agency, which is beyond formalization. Guénin–Carlut [23] warns against the role of the computational metaphor in presenting the results of psychological experiments as laws of behavior, rather than contextual effects. In this view, the brain does not *compute*, but *regulates* its own dynamics to maintain biological autonomy; it does not *represent* the world, but *enacts sense-making* or *realizes relevance*. While we do not intend to dismiss the important nuances being drawn between the organization of information in mechanical and biological systems, we suggest here to explore what could be a minimal representation of what (if anything) is specific about biological computation. Specifically, we suggest that the basic principles of the foundational works of the “anti-computational” approach motivate a relational understanding of biological computation.

Biological autonomy and relational systems

As we briefly evoked above, cognitive science has been defined by two competing poles aiming to model intelligence in opposing ways: explicit computations over symbolic representation of the world *à la* Chomsky [9]; or reconstruction of cognitive dynamics from explicit (if stylized) neuronal models *à la* Rosenblatt [34]. To summarize a long and complex history, the latter approach became largely prevalent in the 80s and 90s with the development of backpropagation [35, 12], which largely overcame the power of any prior architecture and could pretend to some modest level of biological realism. This dominance has only strengthened with the development of deep learning, transformers, etc. But in the same period, perhaps because of the foundational debates regarding the very nature of cognition, some authors started questioning the very premise of describing cognition in terms of computational (symbolic or neural) rather than meta-computational, biologically embedded dynamics. It is in this very period that were written the three references which still stand at the core of most, if not all, arguments against the computational research program in biology and cognitive science. Those are the following: Varela [38]’s *Principles of Biological Autonomy*, Varela et al. [39]’s *The Embodied Mind*, and Rosen [32]’s *Life Itself*.

Varela [38]’s argument is simple, if it has shown difficult to systematize. In his view, the key property of life is its ability to (re)produce its own organization – biological autonomy. And in turn, the regulation of the activity of the organism is embedded in this self-productive process. Thus, biological self-production and cognitive engagement must in some sense be dual. On this basis, Varela et al. [39] propose a novel program for the cognitive science: the enactive approach. In their view, the computational approach to mind (whether symbolic or neural) pre-supposes a fixed organization of meaning, or *cognitive representations* of the world on which the brain performs computational operations. But given the self-produced, out-of-equilibrium dynamics of neural population, such objects may not serve as the basis for cognition. Meaning can only be realized in engagement with the world, and can only be (re)produced itself in same process; there can be no structure of computation and representation prior to cognition, as but only autonomous agents “laying down a path in walking” (Chapter 11) as “mind and world arise together in enaction” (p. 176). Thus, cognitive science should center the process of *enaction* in and of itself, by which agent both mobilize and (re)produce their cognitive structure through active engagement with the world.

Rosen [32] also draws from the basic argument that life is defined by its ability to (re)produce itself. However, rather than drawing from the known physics and biology of this process to investigate the conceptual background of cognitive science, he proposes a full-fledged

¹ University of Sussex
Active Inference Institute
Kairos Research
avel@kairos-research.org
ORCID: 0000-0001-8239-7264

mathematical theory aiming to represent this process. To do so, he builds from his earlier work on measurement theory [31] and anticipation [33], focusing on the mathematical properties of the relation between natural systems and the models we can build by measuring them along a specific collection of observables. On this basis, Rosen [32] derives that if a) an organism is a system *closed to efficient causation*, that is where every component is internally produced; and b) a mechanism is a system that accepts a single sufficient model / can be computed by a Turing machine; then it follows that an organism cannot be a mechanism. On this basis, Rosen [32] introduces a formalization of *relational biology*, describing biological organization in terms of the relation of entailment and production between its individual components.

It would be hard to overstate the similarities between the two approaches. Both are premised on the definition of life as a self-constructive system, and a systematic description of living systems based on this property. Both propose a relational understanding of biological organization, where the significant basis of its definition constitutes the relation between metabolic / production processes rather than intrinsic physical states. And both are largely held by contemporary literature to propose a direct negation of the computational / mechanistic approach to mind and life – which some accept [30, 26] and some oppose [1, 28]. However, a close and adequately contextualized reading of these sources would dispel such a categorical reading. Varela et al. [39] warn us about an excessively static, reified understanding of cognitive representation as a prior basis for cognition. Rosen [32] proposes a demonstration that, under a specific set of definitions, life cannot be reduced to a mechanistic representation. Neither explicitly claims that a mechanistic / computational approach to mind and life cannot capture anything of relevance in its mainstream form, nor that it could not be reformed to understand the specificities of biological systems.

Mechanism and computation

On this basis, let us consider what about classical notion of mechanism and computation arguably fails to capture life and cognition. As defined in the New Mechanistic literature [2, 41, 1], a *mechanism* constitutes a natural structure whose systematic description in its specific context of activity provides an explanation for some phenomena of interest. And as defined in the related program of cognitive neuroscience [5, 28], a *computation* is the manipulation of a medium-independent vehicle according to a specific rule – usually, but not necessarily, executed by a mechanism with a specialized role for processing information about the world. Both programs agree on the important explanatory role of *computational mechanisms* in cognitive neuroscience. It is because the brain can in some form or another *encodes* representations of the world and process them in a structured way that we can come to learn about the world and engage with it adaptively, and the description of the mechanisms by which it does that is an important element of the field as a whole – arguably, its defining element [5].

We do not dispute this characterization at a conceptual level. However, it seems important to point out that applying the concepts of encoding, decoding, and computation to the brain implies a deep tension with their usually understood, mathematically well-defined meaning. As Brette [6] pointed out, Shannon [36]’s canonical information theory defines *encoding* as a process by which a message is embedded in a physical channel *via* a pre-defined procedure so as to be *decoded* by a receiver *via* the inverse procedure. Similarly, Turing [37]’s canonical definition of computation relies on the physical im-

plementation of operations defined within a specific logical algebra. This broad characterization would be implemented by the von Neumann [40] architecture, and every other human-built computational system – including analogical and quantum ones. In other words, encoding and computation are defined mathematically as operations producing, manipulating, and retrieving information embedded in a physical medium *in a pre-defined format*. If someone could produce a compelling argument that the brain operates in such a manner, they kept it close to their chest. Thus, philosophical re-definitions which apply these terms to the brain are best understood as analogical in nature, rather than a conceptual reconstruction of well-defined mathematical theory.

Similarly, mechanistic explanation is by construction premised on the description of a *pre-defined structure*. To be clear, mechanistic description does not presuppose the pre-existence of a specific information format as a computational description does. For example, the ciliary system of the inner ear can filter specific frequencies well above the quality factor accessible to passive processes due to its self-organization in autonomous oscillations near a Hopf bifurcation [25, 10]. This statement is entailed by the bare description of its known biophysical structure, without any assumptions about the information it may carry, manipulate, or transport. Because of this bare-bones nature, such an explanation of the “computational” capacities of the inner ear can hardly be construed as suspect. But for the very same reasons, it does not tell us much about how an organism makes sense of the signal it retrieves from it, and much less how it may reconstruct this meaning through engagement with the world. However useful it may be in describing and explaining biological activity, a small system of differential equations is not “laying down a path in walking”.

Under Varela et al. [39]’s approach to cognition, this relational construction of meaning is precisely the core reason why autonomous systems escape the computational approach. And it is grounded in their ability for the open-ended but structured self-production of the underlying architecture, as described in Varela [38]. In other words, the enactive approach precisely places the *production of organization* – either biochemical or cognitive – at the center of their account of life and mind. Similarly, Rosen [32] constructs life as a kind of organisation which escapes entailment by a prior formal model, thus escaping the domain of mechanistic or computational representation. At first sight, this seems antagonistic to the very premise of a naturalistic approach to describing and explaining the organization of life and mind. How can we account for anything that a living, cognitive agent does, except by representing the underlying structure? However, both approaches center the relational construction of *meaning* [39] and *observation* [32] as the key element which takes autonomous systems beyond the realm of mechanism and computation. On this basis we may ask: is there any manner by which we may account for this relational process in the context of a broadly naturalistic worldview?

Contextual observation for biological systems

To answer this question, we need to take a closer look at Rosen [31, 33, 32]’s specific treatment of the organization of information in natural systems. Rosen [31] in particular provides the formal foundation on which his later work was built, or in the terms of its title the *Fundamentals of Measurement and Representation of Natural Systems*. In this book, he provides a definition of a “measurement” as a physical interaction where a measuring system reaches an indexed state as a function of the internal states of a measured system.

On this basis, he derives various theorems defining the limitations of the positive knowledge we may accrue about the organization of a natural system given any collection of measurements. Most notably, two systems whose dynamics are fully predictable given two respective collections of measurement may present dynamics irreducible to the union of those collections once brought into interaction (Section 4.5). Hence, a description in terms of well-defined dynamics in a well-defined space cannot be derived from observation alone, but requires additional assumptions about its structure – or rather, its lack of structure beyond a given collection of observables. This corresponds to his later characterization of a *mechanism* in [32], which he shows to be in conflict with biological self-production (formalized as closure to efficient causation).

This approach allows us to formalize the intuitive difference between the nature of information in mechanical and biological systems. Mechanisms are intentionally designed to realize a specific formal model, a *blueprint*, and any deviation from the dynamics prescribed by its blueprint is by definition a failure of the realization. In particular, computers are designed as *information-processing* mechanisms: the dynamics their hardware implements are defined so as to correspond to a logical algebra prescribed by their software. In other words, their syntax is by construction bound to implement a specific, pre-defined semantics – as prescribed by formal computational theory. On the contrary, organisms are defined by their self-productive organization, so that their mechanistic structure and/or the information they process are continuously reshaped by the ebb and flow of metabolic processes and biological fluctuations. Their semantics, by construction, are bound to be irreducibly embedded with their syntax – thus (re)producing each other in an open-ended, mutually constructive way. To capture this distinction, we will say that organisms are *autonomous* (etymologically: self-ruled) while mechanisms are *heteronomous* (etymologically: ruled by others).

While autonomy precludes by construction reduction to a single mechanistic model, nothing prevents us from proposing a formal characterization of what is specific about autonomy. In particular, we suggest that the self-production of structure within a biological organization can be captured by capturing the emergence of novel observables as causal variables against which it regulates its activity. The notion of an observation *bringing about* observable states as causal variables, rather than simply *revealing* them, is mathematically well-studied. This property is called *contextuality* [22], and it constituted the most confusing aspect of quantum physics for its early theorists as it suggested a structural incompatibility with a causal, realist model of its object – a suggestion later demonstrated by Bell [3, 4] and Kochen & Specker [27]. Formally speaking, it simply reduces to non-commutativity between evolution operators associated with a collection of observations, so that the resulting state of the system irreducibly and observably depends on the order in which this collection may be applied. But critically, non-commutativity need not be a specific property of atom-scale systems. It may very well be adopted as the model of what happens when an organism “decides” to regulate its activity against a given observable.

Importantly, Fields [13] and Fields & Glazebrook [16] demonstrated that contextuality emerges in any repeated process of observation where the thermodynamic cost of system identification is non-zero. In principle, this applies to biological and mechanical processes alike – including computation. However, Turing machines are structurally constrained to observe the data contained in the specific subsystem which implements their memory – or else fail to fulfill their function. On the contrary, biological systems function on the basis of information embedded in complex, multi-scale organization span-

ning virtually all accessible degrees of freedom [18]. Thus, the system identification cost during the computational process itself cannot be abstracted, and breakdown of commutativity constitutes a systematic property of biological computation across scales. This simple observation affords a rich and physically grounded perspective on the process underlying biological self-production. Biological organization can be framed as a hierarchy (or an heterarchy) of observation processes, all bringing about specific physical states as semantically loaded variables against which the organization regulates its activity [17, 15, 14, 24]. This affords an explicit model of the (re)production of biological structure through the relational construction of meaning, framed in terms of structured information processing – in a broad conceptual sense, *computation*.

Conclusion

In this article, we have questioned the opposition between “computational” [1, 28] and “anti-computational” [30, 26] approaches to life and mind. This is not because we hold the foundational brain-computer analogy to be proportional, or because we consider that Varela [38]; Varela et al. [39]; and Rosen [32]’s arguments failed to establish the deep chasm between the domain of mechanical computation and this of biological and cognitive organization. On the contrary, we consider those arguments to provide a strong demonstration of this case, while also pointing to a specific description of the specific aspect of mind and life which many contemporary authors would recognize as “computational”. The key distinction to be made is that between *heteronomous* organization (intentionally designed to implement a specific formal model) and *autonomous* organization (self-produced beyond entailment by a specific formal model). We suggest the latter are characterized by the production of novel observables as relevant causal variables to their own regulation, through a process of observation embedded in their own biophysical organization. This process explains how organisms may evolve in an open-ended manner, beyond the capacity for representation by any given formal model, while also being legible as a fairly well-known physical property: *contextuality* [22].

To be clear, contextuality does not entail computation in the well-defined formal sense of the term. To enact a contextual observation is to *bring about* the aspects of the world one is observing as significant in one’s own (biological or cognitive) organization, not to *reveal* a pre-existent state of the world or *perform computation* over them within a given logical algebra. For example, the reader can only make sense of this text because they lend meaning to the strange little patterns on their paper or screen by reading them as letters, and the assembly of those letters as words. Yet we know this meaning is not an intrinsic property of those patterns, because we have very direct historical evidence of the cultural evolution of this interpretation in the history of the Latin alphabet and English language. What “contextuality” models is the production of this meaning, and the corresponding evolution of the cognitive / biological structure defining the activity of autonomous agents – see *e.g.* [20, 21] on the sociocultural aspect of this argument. But although it goes by construction beyond the domain of entailment by a formal model, this property also frames the relational production of meaning as a physical process whose structure can be studied through mathematics, characterized in terms of thermodynamic flows and entropy, *i.e.* made legible to a naturalistic program. We are thus confident this proposal will prove itself to be a good compromise, in the limited sense that it will leave no party to the controversy satisfied.

REFERENCES

- [1] William Bechtel, *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*, Taylor & Francis, New York, 2008.
- [2] William Bechtel and Adele Abrahamsen, 'Explanation: A mechanist alternative', *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, **36**(2), 421–441, (June 2005).
- [3] John S. Bell, 'On the Einstein Podolsky Rosen paradox', *Physique Physique Fizika*, **1**(3), (1964).
- [4] John S. Bell, 'On the Problem of Hidden Variables in Quantum Mechanics', *Reviews of Modern Physics*, **38**(3), 447–452, (July 1966).
- [5] Worth Boone and Gualtiero Piccinini, 'The cognitive neuroscience revolution', *Synthese*, **193**(5), 1509–1534, (May 2016).
- [6] Romain Brette, 'Is coding a relevant metaphor for the brain?', *Behavioral and Brain Sciences*, **42**, e215, (January 2019).
- [7] Rodney A. Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**(1), 139–159, (January 1991).
- [8] Rodney A. Brooks, 'New Approaches to Robotics', *Science*, **253**(5025), 1227–1232, (September 1991).
- [9] Noam Chomsky, 'On certain formal properties of grammars', *Information and Control*, **2**(2), 137–167, (June 1959).
- [10] Qian Deng, Fatemeh Ahmadi, William E. Brownell, and Pradeep Sharma, 'The collusion of flexoelectricity and Hopf bifurcation in the hearing mechanism', *Journal of the Mechanics and Physics of Solids*, **130**, 245–261, (September 2019).
- [11] Jean-Pierre Dupuy, *Aux origines des sciences cognitives*, La Découverte, 1999.
- [12] Jeffrey L. Elman, 'Finding Structure in Time', *Cognitive Science*, **14**(2), 179–211, (1990).
- [13] Chris Fields, 'Some Consequences of the Thermodynamic Cost of System Identification', *Entropy*, **20**(10), 797, (October 2018).
- [14] Chris Fields, Filippo Fabrocini, Karl Friston, James F. Glazebrook, Hananel Hazan, Michael Levin, and Antonino Marciandò. Control flow in active inference systems, February 2023.
- [15] Chris Fields, Karl Friston, James F. Glazebrook, and Michael Levin. A free energy principle for generic quantum systems, December 2021.
- [16] Chris Fields and James F. Glazebrook, 'Representing Measurement as a Thermodynamic Symmetry Breaking', *Symmetry*, **12**(5), 810, (May 2020).
- [17] Chris Fields and Michael Levin, 'How do Living Systems Create Meaning?', *Philosophies*, **5**(4), 36, (December 2020).
- [18] Chris Fields and Michael Levin, 'Metabolic limits on classical information processing by biological cells', (arXiv:2103.17061), (March 2021).
- [19] Jerry A. Fodor, *The Language of Thought*, Harvard University Press, Cambridge, 1975.
- [20] Avel Guénin-Carlut. From the existential stance to social constraints - How the human mind becomes embedded in our social, cultural and material context, July 2024.
- [21] Avel Guénin-Carlut. Cognitive agency in socio-cultural evolution - A proposal to ground social modelling in the active inference framework, December 2025.
- [22] Avel Guénin-Carlut, 'Contextuality, Cognitive Engagement, and Active Inference', in *Active Inference*, eds., Christopher L. Buckley, Daniela Cialfi, Pablo Lanillos, Riddhi J. Pitliya, Noor Sajid, Hideaki Shimazaki, Tim Verbelen, and Martijn Wisse, pp. 260–274, Cham, (2025). Springer Nature Switzerland.
- [23] Avel Guénin-Carlut, 'Cognition in eco, cognition in vitro: Measurement, explanation, and contextuality in cognitive sciences', *Intellectica*, 89–116, (January 2026).
- [24] Avel Guénin-Carlut, 'Development is (embodied) intelligence - Biological information, agency, and contextuality', *IOP Conference Series: Materials Science and Engineering*, **1343**(1), 012008, (March 2026).
- [25] A. J. Hudspeth, Frank Jülicher, and Pascal Martin, 'A Critique of the Critical Cochlea: Hopf—a Bifurcation—Is Better Than None', *Journal of Neurophysiology*, **104**(3), 1219–1229, (June 2010).
- [26] Johannes Jaeger, Anna Riedl, Alex Djedovic, John Vervaeke, and Denis Walsh, 'Naturalizing relevance realization: Why agency and cognition are fundamentally not computational', *Frontiers in Psychology*, **15**, (June 2024).
- [27] S. Kochen and E. P. Specker, 'The problem of hidden variables in quantum mechanics', *J. Math. Mech.*, **17**, 59–87, (1967).
- [28] Gualtiero Piccinini, *Neurocognitive Mechanisms: Explaining Biological Cognition*, Oxford University Press, 2020.
- [29] Gualtiero Piccinini, *Neurocognitive Foundations of Mind*, Routledge, September 2025.
- [30] Andrea Roli, Johannes Jaeger, and Stuart A. Kauffman, 'How Organisms Come to Know the World: Fundamental Limits on Artificial General Intelligence', *Frontiers in Ecology and Evolution*, **9**, 806283, (2022).
- [31] Robert Rosen, *Fundamentals of Measurement and Representation of Natural Systems*, North-Holland, 1978.
- [32] Robert Rosen, *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*, Columbia University Press, 1991.
- [33] Robert Rosen, *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*, Springer New York, [2014] 1985.
- [34] F. Rosenblatt, 'The perceptron: A probabilistic model for information storage and organization in the brain', *Psychological Review*, **65**(6), 386–408, (1958).
- [35] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, 'Learning representations by back-propagating errors', *Nature*, **323**(6088), 533–536, (October 1986).
- [36] C. E. Shannon, 'A mathematical theory of communication', *The Bell System Technical Journal*, **27**(3), 379–423, (July 1948).
- [37] Alan Mathison Turing, 'On computable numbers, with an application to the Entscheidungsproblem', *J. of Math.*, **42**(1), 230–265, (1936).
- [38] Francisco J. Varela, *Principles of Biological Autonomy*, North Holland, 1979.
- [39] Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, 1991 (2016 edition).
- [40] J. von Neumann, 'First draft of a report on the EDVAC', *IEEE Annals of the History of Computing*, **15**(4), 27–75, (1945 (1993)).
- [41] Cory Wright and William Bechtel, 'Mechanisms and psychological explanation', in *Philosophy of Psychology and Cognitive Science*, ed., Paul Thagard, Handbook of the Philosophy of Science, 31–79, North Holland, Amsterdam, (January 2007).

The Principle of Excessive Variety: The Brain as an Intrinsically Exploratory System

Kathryn Nave¹

If the point of a good machine is to produce stable, reproducible operations, then the brain is a strikingly poor machine. As much as 40 per cent of its ongoing energy costs are directed towards just maintaining its structure, compared to standard computational components, which can retain information for years with no energy input required (Engl & Attwell D, 2015). Moreover, the brain's stability is not only expensive but also limited. Key synaptic structures regularly fluctuate or disappear entirely independently of any informational input, with typical half-lives in the order of days or weeks (Ziv & Brenner, 2017; Attardo et al. 2015). Over the past decade, the discovery of this pervasive drift has caused consternation among neuroscientists expecting a stable neural code (Driscoll et al., 2017).

But while reproducibility, regularity, and efficiency may be the concern of the scientist, perhaps the brain has something else in mind. After all, intrinsic instability is the default state of a cell, wherein change is cheap and stability must be paid for. And whether or not brains are computers, they are first and foremost networks of living cells. Would they be better if they weren't? Perhaps we should expect artificial networks, which recapitulate some of the computational structure of the brain in stable substrates, to be more effective cognisers in virtue of not having to pay these ongoing costs of organic instantiation. Disconnect these networks from interaction with biological systems entirely, however, and their generative capacity collapses (Shumailov et al., 2024). Perhaps our organic precarity and our capacity for ongoing learning or creative exploration are connected. Various evolutionary and developmental biologists have argued as much (Montévil et al., 2016; Wagner, 2005; Erwin, 2026). Applying their work on the role of random drift in neutral spaces to the phenomena of neural instability helps recast the latter as not an imperfection of a computational brain but an essential characteristic of an exploratory one.

REFERENCES

- [1] E. Engl and D. Attwell, 'Non-signalling Energy Use in the Brain', *The Journal of Physiology* 593 (16), 3417–3429 (2015).
- [2] N. E. Ziv and N. Brenner, 'Synaptic Tenacity or Lack Thereof: Spontaneous Remodeling of Synapses', *Trends in Neurosciences* 41 (2), 89–99 (2018).
- [3] A. Attardo, J. E. Fitzgerald, and M. J. Schnitzer, 'Impermanence of Dendritic Spines in Live Adult CA1 Hippocampus', *Nature* 523, 592–596 (2015).

[4] L. N. Driscoll, N. L. Pettit, M. Minderer, S. N. Chettih, and C. D. Harvey, 'Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex', *Cell* 170 (5), 986–999.e16 (2017).

[5] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, 'AI Models Collapse When Trained on Recursively Generated Data', *Nature* 631, 755–759 (2024).

[6] M. Montévil, M. Mossio, A. Pocheville, and G. Longo, 'Theoretical Principles for Biology: Variation', *Progress in Biophysics and Molecular Biology* 122 (1), 36–50 (2016).

[7] A. Wagner, 'Robustness, Evolvability, and Neutrality', *FEBS Letters* 579 (8), 1772–1778 (2005).

[8] D. H. Erwin, 'A Conceptual Framework of Evolutionary Novelty and Innovation', *Biological Reviews* 96 (1), 1–15 (2021).

¹ University of Edinburgh, Scotland, UK
Kate.Nave@ed.ac.uk